JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF INFORMATION SCIENCE

DEVELOPING A CASE BASED CREDIT APPROVAL SYSTEM USING DATA MINING:

THE CASE OF COMMERCIAL BANK OF ETHIOPIA

By:

WENDWESEN ENDALE

June, 2016

Jimma, Ethiopia

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF INFORMATION SCIENCE

DEVELOPING A CASE BASED CREDIT APPROVAL SYSTEM USING DATA MINING:

THE CASE OF COMMERCIAL BANK OF ETHIOPIA

A Thesis Submitted in Partial Fulfillment of the Requirements for Degree of Masters of Science in Information Science (Information and Knowledge Management)

By:
WENDWESEN ENDALE

Principal Advisor: Million Meshesha (PhD)

Co-Advisor: Amanuel Ayde (MSc)

June, 2016
Jimma, Ethiopia

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF INFORMATION SCIENCE

DEVELOPING A CASE BASED CREDIT APPROVAL SYSTEM USING DATA MINING:

THE CASE OF COMMERCIAL BANK OF ETHIOPIA

By:

WENDWESEN ENDALE

As members of the board of examining of the MSc thesis open defense examination of the above title, we members of the board (listed below), read and evaluated the thesis and examined the candidate.

Name	Title	Signature	Date
Yeshihareg Temteme	Chairperson	______________	__________________
Million Meshesha (PhD)	Principal Advisor	______________	__________________
Amanuel Ayde (MSc)	Co-Advisor	______________	__________________
Tibebe Besha (PhD)	External Examiner	______________	__________________
Elsabet Wedajo (Msc)	Internal Examiner	______________	__________________

DECLARATION

I declare that this thesis is my original work and it has not been presented for a degree in any other universities. All the material sources used in this work are duly acknowledged.

____

Wendwesen Endale

June, 2016

This thesis has been submitted to the department for examination with our approval as university advisors:

Principal Advisor: Million Meshesha (PhD) …………………….

Co-Advisor: Amauel Ayde (MSc) …………………….

June, 2016

DEDICATION

This work is dedicated to my lovely wife Hiwot Asefa.

ACKNOWLEDGEMENT

First of all, I would like to gratitude almighty God for giving me strength and wisdom to complete this thesis work. Next to God, I would like to take this opportunity to express my profound gratitude and deep regard to my adviser, Dr. Million Meshesha and Ato Amanuel Ayde, for their exemplary guidance, valuable feedback and constant encouragement throughout the duration of the research. Their valuable suggestions were of immense help throughout the research.

I express my warm thanks to Mettu University, Jimma University Department of Information Science staffs , my families and all the friends and colleagues who provided me with the facilities being required and conductive conditions for this research.

I would also like to thank you CBE head office and Jimma District staffs for their kindly help and support. Especially Mr. Henoke this research would not be successful without your help and encouragement.

Finally, I would also like to give my sincere gratitude to my beloved wife, Hiwot Assefa , for her encouragement, support, challenging comments and being by my side without whom this research would be incomplete.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
ARFF	Attribute-Relation File Format
CBE	Commercial Bank of Ethiopia
CBR	Case based Reasoning
CBRCADM	Case Based Reasoning system for Credit Approval Decision Making
CRM	Credit Risk Management
CSV	Comma Delimited value File
DM	Data Mining
FI	Financial Institutions
ES:	Expert System
JCOLIBRI	Java Case and Ontology Libraries Integration for Building Reasoning infrastructures
KBS	Knowledge Based Systems
KDD	Knowledge Discovery in Database
RBS	Rule Based Reasoning
NBE	National Bank of Ethiopia
WEKA	Waikato Environment for Knowledge Analysis
XML	Extended-Markup Language

Abstract

The very nature of the banking business is so sensitive because more than 85% of their liability is deposits from customers. Banks use these deposits to generate credit for their borrowers, which in fact is a revenue generating activity for most banks. However, this credit creation process exposes Banks to high default risk, which might lead to financial distress including bankruptcy. So banks need to manage credit risk inherent in the entire portfolio as well as the risk in individual credits or transactions. In this study therefore, a case-based credit approval decision making knowledge base system that uses data mining results is proposed by applying empirical research design. The researcher used manual and automated knowledge acquisition techniques, such as interview, document analysis and data mining. To identify the best prediction model for Credit approval decision making, three experiments using three classification algorithms were conducted. Finally, the researcher decided to use the results of J48 decision tree classification algorithm in the development of the prototype case-base System because it registered better performance than other classifiers. The developed model was tested with test instances and only those instances registers more than 95% accuracy were used to develop a knowledge base for the CBR development for a better efficiency. Then, the implementation of the prototype using JCOLIBERI version 1.1 which is object oriented case-based reasoning framework is realized. Finally, testing of the prototype case-based reasoning system is done to evaluate the performance of the system. The prototype is evaluated using system testing and user acceptance testing. Testing system performance in terms of precision, recall and f-measure registered 83%, 73 % and 77 %, respectively. Also user acceptance testing achieved 83.2% performance. The evaluation of the prototype shows a promising result to design an applicable intelligent system that supports effective and efficient credit approval decisions making. But, the current system suggests no explanation about the correct action to be taken; as a result a hybrid explanation driven system by combining Case based reasoning with Rule based reasoning is recommended as a future research direction.

CHAPTER ONE

INTRODUCTION

1.1. Background of the study

Banks are financial institutions that are established for lending, borrowing, issuing, exchanging, taking deposits, safeguarding or handling money under the laws and guidelines of a respective country. Among their activities, credit provision is the main product which banks provide to potential business entrepreneurs as a main source of generating income (Mirach, 2010).

According to Boru (2014) banks play a vital role in economic development through engaging themselves in an intermediary role which enhances investment and growth. In line with this commercial banks contribute positively to economic growth by channeling surplus funds to their most productive uses. Mirach (2010) defined credit risk as the risk of loss of principal or loss of a financial reward stemming from a borrower's failure to repay a credit or otherwise meet a contractual obligation. The very nature of the banking business is so sensitive because more than 85% of their liability is deposits from depositors (Khac & Kechadi, 2010). Banks use these deposits to generate credit for their borrowers, which in fact is a revenue generating activity for most banks. This credit creation process exposes the Banks to high default risk, which might lead to financial distress including bankruptcy. In today’s changing financial landscape-environment of intense competitive pressure, volatile economic conditions, rising bankruptcies, and increasing levels of consumer and commercial debt; an organization’s ability to effectively monitor and manage its credit risk can mean the difference between success and failure.

Risk is the fundamental element that drives financial behavior. Without risk, the financial system would be vastly simplified. However, risk is omnipresent in the real world. Financial Institutions, therefore, should manage the risk efficiently to survive in this highly uncertain world (Sahlemichael, 2009).

Risk management is a discipline at the core of every financial institution and encompasses all activities that affect its risk profile. It involves identification, measurement, monitoring & controlling risks to ensure that the individuals who take or manage risks clearly understand that the organization’s risk exposure is within the limits established by board of directors. Risk taking decisions are in line with the business strategy and objectives set by board of directors.

Based on Bank Supervision Directorate report (2016), Banks need to manage credit risk inherent in the entire portfolio as well as the risk in individual credits or transactions. Additionally, banks should be aware that credit risk does not exist in isolation from other risks, but is closely intertwined with those risks. Effective credit risk management is the process of managing an institution’s activities which create credit risk exposures, in a manner that significantly reduces the likelihood that such activities will impact negatively on a bank’s earnings and capital. While providing credit as a main source of generating income, banks take into account many considerations as a factor of credit management which helps them to minimize the risk of default that results in financial distress and bankruptcy. This is due to the reason that while banks providing credit they are exposed to risk of default which need to be managed effectively to acquire the required level of credit growth and performance (Mekonnen, 2009).

Effective credit risk management attracts today more attention than before. CBR is an Artificial Intelligence technique used to solve problems where human analogy reasoning is simulated. Through analogy reasoning, new problems are solved by the adaptation of solutions used to solve previous similar problems. The most significant advantage of CBR is that its ability to search, calls up, reuses and adapts past experiences (representing previous specific problems) to solve an existing situation (Aamodt & Plaza, 1994).

CBR method is based on the assumption that, similar problems with similar solutions tend to occur (Lang & Lau, 2002). Different scholars have witnessed a growing number of business applications of CBR (Triki & Bellamine, 2013). CBR allows businesses to treat past cases as a corporate resource, which can be used in future for making decisions. The recent downsizing of several corporations has also increased the need to maintain a corporate memory, because of the large turnover of highly skilled professionals. Domains that are endowed with a rich history of cases are particularly suitable for CBR (Bergmann, *et al.*, 2005). One such domain is bank lending; banks usually have available a large number of past credit applications (cases).

In Case Based Reasoning it is likely that the user will have better confidence in that outcome, but logistic regression, as well as neural nets and rule-based systems, does not give the user guidance on how to proceed and this techniques were not applicable in domain where there are no fully understood and well-structured general rules were available that capture the relationships between problems and credit decisions (Juan, *et al.*, 2009). Those models do not provide exemplars and therefore cannot aid the user in deciding on borderline cases in the same way that CBR does. Based on the empirical study carried out it is possible to state that in credit exposure analysis CBR systems are more accurate than the statistical techniques (Discriminates analysis and Logistic regression) usually applied in this type of decisions (Costa, *et al., 2008).*

What gives a CBR system its real power is, its ability to look back at past experiences, both successes and failures. This type of feedback encourages the system to repeat its past successes, while, at the same time, warns it against potential pitfalls (Triki & Bellamine, 2013).

But there are obvious problems in choosing a single technique for solving problems in poorly understood domains, where modeling is complex or information is difficult to acquire. A variety of techniques have been used to solve the problems that arise during the CBR cycle. There has been increasing development of hybrid systems (Charlo & María, 2010). Since knowledge is incomplete and dynamic, we have to expend our options through which we can acquire knowledge from different sources such that we can make the knowledge base of the Knowledge Based System as complete as possible. In order to make knowledge extraction as much correct as possible different techniques could be applied. According to Amritpal *et al.* (2015) findings, data mining or knowledge discovery techniques became the most used in the recent years. The cornerstone of an effective Knowledge-Based System is data mining.

Data mining improves decision making by giving insight into what is happening today and by helping predict what will happen tomorrow (Amritpal *et al.*, 2015). Data mining is a subfield of Machine Learning that enables finding interesting knowledge (patterns, models and relationships) in very large databases. It is the most essential part of the knowledge-discovery process, which combines databases, statistics, artificial intelligence and machine learning techniques (Bhambri, 2011).

On top of this, Credit is the common fundamental component in any financial system. Different economic units, both demand and supply side, require credits. Individual require credit for economic and social need. Governments also require credit for financing its deficit and building public project. Most of Businesses organizations heavily relay on external source of finance for expansion, modernization, working capital requirement and financing new project. As a result of it, Credit risk is the most important risk banks face specially in developing country like Ethiopia where no formal stock market available as alternative source of external fund (Atakelt & Veni, 2015). As a result, bank service is considered as life blood for any economic unit while effective Credit risk management system and practice is an ingredient part of safety, soundness, liquidity and profitability of banks.

Assuredly, the axle of this research is to mine the bank credit database, classify the credit applicants into different groups, and serve as a tool for establishing better credit approval decision making system by using the result in to the knowledge base system.

1.2. Statement of the problem

The future of banking will undoubtedly rest on risk management dynamics. Only those banks that have efficient risk management system will survive in the market in the long run (Gizawe, *et al*., 2015). The very nature of the banking business is so sensitive because more than 85% of their liability is deposits from depositors (Khac & Kechadi, 2010). Banks use these deposits to generate credit for their borrowers, which in fact is a revenue generating activity for most banks. This credit creation process exposes the banks to high default risk which might led to financial distress including bankruptcy. All the same, beside other services, banks must create credit for their clients to make some money, grow and survive so as to gain competitive advantage at the market place.

Consequently, the effective management of credit risk is a critical component of comprehensive risk management essential for long-term success of a banking institution. Credit risk is the oldest and biggest risk that bank, by virtue of its very nature of business, inherits (Mirach, 2010), (Atakelt & Veni, 2015). This is because credit risk can easily and most likely prompts bank failure (Basle committee on banking supervision, 2004). As stated by Bank Supervision Directorate (2010), Risk-taking is an inherent element of banking and, indeed, profits are the reward for successful risk taking. In line with this, excessive, poorly managed risk can lead to distresses and failures of banks. Risks are, therefore, warranted when they are understandable, measurable, controllable and within a bank's capacity to withstand adverse results.

Banks move in to a new high powered world of financial operations and trading, with new risks, the need is felt for more sophisticated and versatile instruments for risk assessment, monitoring and controlling risk exposures. According Gizawe *et al*.(2015) research, banks that had been performing well suddenly announced large losses due to either credit exposures that turned bad, liquidity problems, or significant operational risks.

As a result of a research conducted by Mekonnen (2009) on Credit Risk Management System of Ethiopian Commercial Bank during the last 8 years period (2000- 2008), the banks acknowledged that they had faced risks in reply for the questionnaire. In addition to this, based on Gizawe *et al.* (2015) findings, credit risk is the most common and frequently occurring risk in the commercial banks and they recommends that , all banks are encouraged to develop and utilize an internal risk rating System to manage credit risk at any time. In response to such lessons, banks almost universally have embarked up on an upgrading of their risk management systems and credit approval mechanisms. According to Bank Supervision Directorate (2010), Credit risk measures have a significant impact on the profitability of commercial banks in Ethiopia and it is suggested that a rigor credit risk management information system is in a paramount importance.

Although underdeveloped, the banking system in Ethiopia has witnessed a significant expansion over the past few years. The National Bank of Ethiopia (NBE) believes such growth should be matched to strong risk management practices. As a result, the NBE has revised the risk management framework it issued in 2003 to all banks so as to incorporate latest developments in the area, which is consistent with international standards and best practices, and expected to provide minimum risk management (risk identification, measurement, monitoring and control) standards for all banks operating in the country (Gizaw, *etal.*, 2015). It is, therefore, time that banks managements equip them fully to go with the demands of creating tools and systems capable of assessing, monitoring and controlling risk exposures in a more scientific manner before happening (Boru, 2014).

Currently CBE used manual credit approval mechanisms. The main tool they used to restrict and control the credit risk taken by the Bank is the credit limit technique. Credit risk limits are determined by the Credit Committee and approved by the Bank's Management Board (Mirach, 2010). Unfortunately such ways can't fully prevent the credit risks in the entire customers portfolio and are not able to give them a versatile support for their decision making purpose since process of making credit evaluation decision is complex and unstructured (Gizawe *et al*.,2015), (Mekonnen, 2009). Especially in the case of banks, the issue of credit management is of even greater concern because of the higher levels of perceived risks resulting from some of the characteristics of clients, business conditions and economic environment in which they find themselves.

Case-Based Reasoning (CBR) is promising to build the decision making system for credit approval and previous experiences have witnessed a growing number of business applications of CBR (Triki & Bellamine, 2013). CBR allows businesses companies to treat past cases as a corporate resource, which can be used in future for making decisions. Domains that are endowed with a rich history of cases are particularly suitable for CBR (Bergmann, *et a*l., 2005). One such domain is bank lending; banks usually have available a large number of past credit applications (cases).

In addition, Data mining is used as a tool in banking and finance in general to discover useful information from the operational and historical data to enable better decision-making and it also helps automated knowledge acquisition mechanism in such poorly understandable domains (Khac *et al.*, 2011), (Kazi & Ahmed, 2012).

Even though many studies have been conducted before using DM or CBR separately for credit, the requirement on technical support for its effective and efficient decisions making is still in high demand (Amritpal, *et al.*, 2015; Bhambri, 2011; Costa, *et al.*, 2007; Ionita & Ionita, 2011).

Maria & Lusi in 2002 used CBR approach with the main objective is to enhance the knowledge of credit standards and credit conditions in the euro area and obtained a promising result to the decision of economic agents for forecasting the behavior of economic agents in the credit condition. Since their work is only considers euro system it can't be a representative for other areas like our own. The current work is also different from Maria & Lusi (2002) since it used the application of data mining for attribute selection and knowledge base construction.

Besides CBR, different scholars also used data mining techniques for Credit Approval which aims to evaluate the performance and accuracy of classification models (Choge ,2012), (Chitra & Subashini ,2013).But the current work is different from the previous research findings with since it provides GUI based help for credit case decision makers by using CBR techniques by uses a combined DM and CBR approach using local data set.

Therefore, in order to facilitate credit approval process and to support credit officers' in their decision making, this study aims to design and develop a Case-based reasoning system using data mining results as a knowledge source. To the end, the study attempt to explore and answer the following research questions.

- What are the main attributes that can properly predict the type of decision given for new credit applicants?
- Which classification algorithm is best to develop the prediction model that can help for credit approval decision making?
- How a CBR prototype system is developed by using DM results as Knowledge source for credit risk decision making?
- What is the CBR system performance and user acceptance of the prototype?

1.3. Objective of the Study

1.3.1. General objective

The general objective of the study is to design and develop a Case Based Reasoning system by using data mining results with case based reasoning that support loan experts for effective and efficient Credit Approval decisions making.

1.3.2. Specific objectives

- To acquire and identify main attributes that can properly predict the type of decision given for new credit applicants.
- To explore and identify suitable classification data mining technique and algorithm for automatic knowledge acquisition.
- To represent and model the acquired knowledge.
- To develop a prototype CBR system that can help for credit approval decision making by using DM results as knowledge source.
- To evaluate the performance and user acceptance of the proposed case based reasoning system.

1.4. Scope and Limitation of the study

The scope of this study is to design and develop a prototype case based system for credit risk decision making at Commercial bank of Ethiopia. The researcher selects CBE based on the assumption that since CBE is the oldest Public Commercial Bank, it has a better credit history and experienced officers than other commercial banks. The knowledge for the case based system was acquired from domain experts' interview, documents analysis and banks credit data set by employing classification data mining techniques. It also discusses the use of case-based reasoning system to solve a domain- specific problem - namely, credit risk management.

The main constraint that the researcher faced while doing this researcher is data size. Since CBE start handling and reporting the customer data in softcopy to NBE Central Data Base after 2007 based on the Licensing and Supervision of banking business directive which comes into effect as of the 15th day of April 2006, the size of the data become only 1518 with 18 attributes after preprocessing. The data set the researcher used are data from September 2007 - August 2012 which are prepared with the aim of reporting the CBE credit customer's data to NBE. In line with this time and budget is another constraint of the study. Due to the above mentioned constraints enough data set cannot obtained since the data is collected from only Head Office of CBE. This is due to the reason that other districts handle their credit customer information in hard copy and they assumed the data is confidential to give it with its all features in hardcopy. As a result some important information's for the decision making process like Customers Historical financial analysis, Appraisal and relationship experts recommendations Liquidity Business plan status, and Collateral audit etc which were kept in hard copy formats are not included in this study.

1.5. Significance of the Study

Credits and advances are known to be the main stay of all commercial banks. They occupy an important part in gross earnings and net profit of banks. The share advances in the total asset of banks forms a lion share (almost more than 60 percent) and as such it is the back bone of banking sector. Bank lending is very crucial for it makes possible the financing of agricultural, industrial, construction, and commercial activities of a country. The strength and soundness of the banking system primarily depends upon health of the advances. Therefore the ability of banks to formulate and adhere to policies, procedures and systems that promote credit quality and curtail non- performing credit is the means to survive in the stiff competition. In ability to create and build up quality credits and credit worthy customers leads to default risk and bankruptcy as well as hampers economic growth of a country.

This study is significant in providing a system based help for credit approval decision making to enhance the performance of credit management to all managers and decision makers in the complex banking environment. Moreover, it helps as a benchmark for researchers who are interested in the area to extend it further. The developed prototype CBR system can be used to give advising services for bank credit officers during credit cases approval.

1.6. Methodology

1.6.1. Research Design

This study follows empirical research design. According to Goodwin (2005), empirical research is research using empirical evidence. It is a way of gaining knowledge by means of direct and indirect experience. As a result, in this study the researcher used experimental method for model building, analysis, and prototype development and testing, whereas non-experimental method was used for knowledge elicitation through discussion with experts and document review.

1.6.2. Study Area

The main data source used for this study is data set of previously solved cases at commercial bank of Ethiopia and domain experts working in the aforementioned organizations. The researcher collected the dataset for data mining purpose from CBE Head Office and data for prototype testing from Jimma district CBE. The Researcher selected this organization with the assumption of, since it is one of the oldest public Commercial Bank in Ethiopia and has a better credit cases and experienced experts than other banks. In order to get the required information for the research and comments at different stage of experimentation and evaluation, discussions and unstructured interview was conducted with purposively selected domain experts at CBE.

1.6.3. Knowledge Acquisition

The researcher used both manual and automated knowledge acquisition mechanisms.

1.6.3.1. Manual Knowledge Acquisition

The researcher used both interview and documents analysis to acquire knowledge. The researcher conducted the domain experts' interview with Customer relationship managers (CRM), Loan officers, and Branch managers who works in CBE from Head office and Jimma district that have working experience in credit approval. The researcher conducted the domain experts' interview with 6 experts which are purposively selected from CBE Jimma district due to easy of access and time limitations to get domain know how about credit approval process and for user acceptance testing of the prototype.

1.6.3.2. Automated Knowledge Acquisition

The researcher used Knowledge Discovery in Database (KDD) process model to automatically acquire knowledge from the CBE dataset using Waikato Environment for Knowledge Analysis (WEKA) version 3.6.5 data mining tool.

KDD is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions. Generally, there are five steps in the KDD process (Two Crows Corporation, 1999; Azevedo & Santos 2008).

Data selection: This stage consists on creating a target dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed. The data relevant to the analysis is decided on and retrieved from the data collection. The researcher selects the data set from CBE head office. The data set the researcher used are data from September 2007- August 2012 which are prepared with the aim of reporting the CBE credit customers data to NBE which have a known final status about their final credit payment.

Data pre-processing: This stage consists on the target data cleaning and pre-processing in order to obtain consistent data. The researcher used dataset from a flat file or a spreadsheet.

Since the researcher uses a dataset which is preprocessed for the sake of reporting the customer credit reports to NBE central credit database which is in a flat file or a spreadsheet format, it passes most of the pre-processing steps. But the researcher performs other preprocessing activities to make the data more suitable for data mining like data cleaning, removing attributes and handling missing values.

Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure. This stage consists on the transformation of the data using dimensionality reduction or transformation methods. As a result makes data transformation for some selected attributes to make the data more suitable for data mining within different ranges.

Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns. It consists on the searching for patterns of interest in a particular representational form, depending on the DM objective. The researcher used classification technique on Bank data set which have been collected from CBE to develop a model that can predict the credibility status of the customer so that to use the model for case based development. Classification is form of data analysis that can be used to extract models describing important data classes or to predict future data trends and classification predicts categorical (discrete, unordered) labels (Asghar & Iqbal, 2009).

The researcher conducted three experiments for three classification algorithms namelyJ48 pruned, PART and naïve Bayes.

For conducting this research the WEKA (Waikato Environment for Knowledge Analysis) version 3.6.5 (for windows OS) DM software is chosen. WEKA is chosen based on DM tool selection criteria of Collier et al. (1999) in addition to its widespread application in different DM researches and familiarity of the researcher with the software. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality (Khac & Kechadi, 2010). It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code. The data provides any meaningful information that can be used to know anything about any object (Ionita & Ionita, 2011).

As noted by Witten & Frank, (2000), advantages of Weka include the following:

- Free availability under the GNU General Public License
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
- A comprehensive collection of data pre-processing and modeling techniques
- Ease of use

Interpretation/Evaluation: This stage consists on the interpretation and evaluation of the mined patterns. Model creation is followed by performance evaluation which measures the accuracy rate of the system. The mined pattern enables to identify the truly interesting ones. For any errors or mismatched result generation as compared to domain area perspectives, the process restarts to initial step so as to provide accurate results. In DM evaluation serves two purposes. First, it helps to envisage how well the final model will work in the future (or even whether it should be used at all). Second, as an integral part of many learning methods, it helps to explore the model that best represents the training data. Accuracy means the percentage of test set samples that are correctly classified by the classifier.

Finally, visualization and Knowledge representation are used to present the mined Knowledge to the users and stored as new Knowledge in the Knowledge base. Incorporating the Knowledge in to another system for implementation purpose, documentation and report for presenting the benefit of the Knowledge to interested parties, incorporating the Knowledge with previously Known Knowledge in the area are some of the important activities during this phase. Likewise, classification models that are developed in this research are evaluated using a test dataset based on their classification accuracy and interpretation also made accordingly. As a result a test instance which registers more than 95 % accuracy was taken as a knowledge source for CBR development.

1.6.4. Case Representation

In the process of case based reasoning system development, case representation is one of the basic steps. It refers to the formalism, both syntax and semantics, used to store knowledge in the architecture (Bergmann, *et al.*, 2005). It is also the process of interpreting domain knowledge into computer understandable form using various knowledge representation techniques. The object of a knowledge representation is to express knowledge in a computer tractable form, so that it can be used to enable our AI agents to perform well (Birmingham & Klinker, 2009).

The common Knowledge representation techniques include semantic network, logics, rules, case base and frames (Inderpal, 2013). Among these, the researcher uses case based representation method for this research. The choice of a CBR representation is linked to the fact that the technique provides automatic knowledge acquisition in poorly understood areas which are not easily approached through other techniques and CBR can even function as a cognitive model to forecast and analyze aspects of human thought and behavior using past experiences (Aamodt & Plaza, 1994).

For this research the researcher represented the knowledge that comes from the manual knowledge acquisition mechanisms through conceptual modeling, data mining results as rules and used feature-value case representation for case based development. The reason for representing the cases using feature-value representation is that this approach supports nearest neighbor retrieval algorithm and it represents cases in an easy way (Salem, 2005), (Ethiopia, 2002).This approach also uses old experiences to understand and solve new problems. It also reuses its solutions and lessons learned for future use. In addition, it represents cases in an easy way by using attribute and value pair representation (Birmingham & Klinker, 2009). Case retrieval (similarity measurement) usually falls into one of four categories: nearest neighbor, inductive learning, knowledge-guide, and a combination of these (Buta, 1994). The algorithm used to calculate the similarity of cases in a case base representation for this research is nearest neighbor retrieval algorithm. The similarity function of nearest neighbor retrieval algorithm involves in computing the similarity between the stored cases in the case base and the new query. After that, it selects the most similar stored cases to the query. . It is suitable when there are attributes that have numeric (continuous) value (Fag & Songdong, 2007).

1.6.5 System development methodology

Prototyping approach is followed to develop the case based system. Prototyping allows participating users and domain experts for evaluating systems performance and efficiency.

1.6.6. Implementation tools

To develop Case based systems there are various tools which are available both freely and commercially. Among this myCBR, and JCOLIBRI are among the most widely used and known frameworks for teaching and academic research purpose (Antanassov & Antonov, 2012). Both of the aforementioned tools have their own capabilities and limitations. A CBR tool could be used to develop several applications that require case based reasoning methodology. Hence in this study for the development of CBR prototype system, the researcher used JCOLIBERI version 1.1 which is object oriented framework and data mining tool WEKA version 3.6.5. WEKA is chosen since it is proven to be powerful for data mining and used by many researchers for mining task and the researcher is familiar with the tool. It contains tools for data preprocessing, clustering, regression, classification, association rules and visualization.

According to Triki & Bellamine (2013), the major advantage of JCOLIBERI comes from its support of full CBR cycle (retrieve, reuse, revise, retain). It is also suitable for developing large scale applications works well with external data base, extensible framework and compatible with different applications as its developed based on object oriented framework.

According to Juan, et al. (2009), JCOLIBRI framework is also extensible, reusable and can be used with different types of users and different purposes (development, research and/or teaching), compatible with commercial applications and, supporting different types of CBR systems, since it is just a .jar file suitable for web applications. It is suitable for developing large scale applications and it works well in external database.

1.6.7. Evaluation methods

Once the prototype is developed, the functionality and user acceptance of the system should be tested. The evaluation processes focus on system's user acceptance of the prototype and the performance of the system.

The researcher used Precision, Recall, F-measure and True Positive rate to evaluate the results and accuracy of the data mining model. The researcher also evaluated the prototype CBR using system performance testing using Precision, Recall, F-measure by preparing test cases and users' acceptance testing questionnaire which helps the researcher to make sure that whether the potential users would like to use the proposed system and whether the proposed systems meets user requirements

.

1.7. Operational definitions

- **Credit risk**: is the risk of loss of principal or loss of a financial reward stemming from a borrower's failure to repay a loan or otherwise meet a contractual obligation.
- **Risk Management**: the identification, analysis, control, minimization or elimination of unaccepted risks.
- **Combining:** refers to the end result of one system will be used as input for the other sub system.
- **Domain Expert**: - is a person who expertise in his/her domain area. In addition, a network administrator who manages and administers a given network is a domain expert in his domain.
- **Knowledge Engineer**: - is one who gathers knowledge from experts through interview or using automatic knowledge acquisition techniques. The knowledge engineer has to have the knowledge of a knowledge base development technology and should know how to develop Case based system using a development environment.

1.8. Organization of the study

This study comprises six chapters. Chapter one discusses background of the study, the problem statement and research questions, the general and the specific objectives of the study, and methodologies that the researcher used to conduct this study.

Chapter two discusses about conceptual and related works review that are relevant for this study. In this chapter, the researcher discussions about Case Based System including CBR cycles, CBR System Performance Evaluation Methods, Data Mining and knowledge discovery concepts, Application areas of Data Mining in banking, DM models and related works which are relevant for this study.

Chapter three presents the knowledge acquisition process. The focus here is on manual (domain expert interview and documents analysis) and automated knowledge acquisition techniques through data mining. After the manual knowledge acquisition step, the researcher proposed the conceptual model for credit approval decision making. The researcher presented the Knowledge discovery steps such as data set preparation, preprocessing, predictive model creation and experimentation.

The researcher also discussed the results of WEKA classifier algorithms by comparing one to another.

Chapter four discusses about Design and Implementation. In this chapter the design and implementation of the prototype are realized by using Data mining results as a knowledge source. The architecture of the new prototype CBR system for Credit Approval decision making is developed. The implementation tool used is JCOLIBRI version 1.1.

Chapter five discusses about implementation and evaluation of the proposed systems. In this chapter the performance of the prototype is evaluated both the performance of the system and the acceptance of the system by the users. In addition, discussion was made to show the significance of the proposed approach with previous researches.

Finally, the researcher dedicated chapter six for conclusion and recommendation. In this chapter, the researcher discussed the evaluation results and based on the result the researcher presents findings and concludes the study by recommending future works.

CHAPTER TWO

LITRATURE REVIEW

The aspiration for computer systems being able to support human experts during complex problem-solving task is a usual topic of AI research. In order to enable a computer system to give rational support when solving problems in a complex application domain, it is essential to provide it with specific knowledge within that domain. A number of methodologies to realize such knowledge knowledge-based systems have been developed, such as, rule-based approach. In recent years, CBR has become a very popular technique for developing knowledge-based systems that can give rational support using specific knowledge. In some real-world application domains, it has even emerged to one of the commercially most successful approaches compared to other techniques developed in AI research such as rule-based systems. Using CBR, different applications have been developed yet, to solve problems in different domains. CBR is useful for a wide variety of problem-solving tasks like planning, diagnosis, design and decision support (Kolodner, 1992; Lenz, et al, 1998). In order to have deep understanding on the problem of this study, it is vital to review several literatures that have been conducted in the field so far. For this reason, related literature such as books, journal articles, proceeding papers, magazines, manuals and some other sources that are retrieved from the internet have been consulted so as to understand the domain knowledge, concepts, principles and methods that are important for achieving the research objective.

2.1. Case-based Reasoning

CBR is a conventional of new theory and research method developed in the domain of Artificial Intelligence (AI) in 1977 by Schank and Abelson. According to Aamodt & Plaza (1994) CBR is an AI technique to support the capability of reasoning and learning in advanced decision support systems. Specifically, it is a reasoning paradigm that exploits the specific knowledge collected on previously encountered and solved situations, which are known as cases. According to Pal & Shiu (2004), CBR means reasoning from experiences or old cases in an effort to solve problems, critique solutions, and explaining inconsistent situations. Instead of modeling a complete domain theory, for example, by using rules, CBR exploits single situation-specific knowledge chunks called cases, which are easier available than generalized knowledge about the domain. The CBR method grew rapidly over the last few years, as seen by its increased applications and in terms of papers at major seminars, available commercial tools, and applications in daily use.

A case-based reasoned will be presented with a problem, either by a user or by a program or system (Pal & Shiu, 2004). The case-based reasoner then searches its memory of past cases (called the case base) and attempts to find a case that has the same problem specification as the case under analysis. If the reasoner is not able to find an identical case in its case base, it will attempt to find other case(s) that are close to match the current case.

In situations where a previous identical case is retrieved, assuming that its solution was successful, it can be accessible as a solution to the current problem at hand. In the more likely situation that the case retrieved is not identical to the current case, an adaptation phase occurs. During adaptation phase, differences between the current and retrieved cases are first identified and then the solution associated with the case retrieved is modified, taking these differences into account. The solution returned in line with the current problem specification may then be tried gain after the suitable domain setting.

2.2. CBR Cycle

At the heart of the system, CBR has been formalized for purposes of computer reasoning as a four-step process which are mostly called 4REs (Retrieval, Reuse, Revise and Retain) (Aamodt & Plaza, 1994).

2.2.1. Retrieval

In CBR, retrieval is remembering previous cases stored in the case base to solve new problems at hand. The first step which is the most important step in CBR cycle is retrieval of previous cases that can be used to solve the target situation (new problem). Given a target problem, retrieve cases from memory those are relevant to solving it. A case consists of a problem description, its solution, and typically annotations about how the solution was derived (Aamodt & Plaza, 1994). Since retrieval is the first step in CBR, it affects the whole CBR system because others cycles are based on it. The selection of corresponding useful cases is then left to the CBR system which retrieves cases to be used for solving the problem by employing so called similarity measures. To retrieve relevant cases to the target problem, appropriate similarity measurement should be used

(Mantaras *et al*., 2005).

According to the basic CBR assumption (similar problems have similar solutions) and here the concept of similarity is used. This means, the task of the retrieval phase is to select cases whose problem descriptions are similar to the current problem's description. The underlying assumption is that these cases contain solutions being very similar to the searched, but still unknown solution of the current problem (Aamodt & Plaza, 1994; Mantaras *et al.*, 2005).

To realize this retrieval task, CBR systems employ special similarity measures that allow the computation of the similarity between two problem descriptions. Because the interpretation of this similarity strongly depends on the particular domain, similarity measures are part of the general knowledge of the system.

Different researchers describe the main tasks during retrieval of cases. For example - Aamodt & Plaza (1994) grouped case retrieval subtasks into three:

- **Identify features**. Involves indexing the problem with the most descriptive feature in order to match it with indexed matched cases. In other words, it identifies its descriptive properties and takes out the properties which don't describe the problem strongly.
- **Initially match**. Finding previous cases that match with the problem at hand and it retrieves a set of plausible candidates. That means it involves searching and similarity assessment to produce similar cases.
- **Select.** Selecting the best-matched case from the set of similar cases. It is based on the similarity assessment result that best matched case or set of cases is selected as output of the retrieval process.

The quality of the retrieval process depends on its descriptive feature identifying algorithm, searching algorithm and similarity assessment method. In CBR there are different case retrieval algorithms but the two most frequently used are nearest neighbor and induction case retrieval algorithms (Singh *et al.*, 2007). These algorithms can be used alone or in combination with each other.

Nearest Neighbor Retrieval Algorithm

Nearest-neighbor retrieval technique is to measure similarity between the source case and the case which we are searching (Lang & Lau, 2002). The nearest neighbor algorithm measures the similarity of stored cases with a new input case, based on matching a weighted sum of features (Watson & Marir, 1994; Singh *et al.*, 2007). When a new case doesn't exactly match with old cases then this algorithm will return nearest match from CBR library. It is suitable when there are attributes that have numeric (continuous) value (Fag & Songdong, 2007). But the retrieval time of this algorithm increases linearly as the case in the case base increases.

The algorithm for nearest neighbor is as follows (Salem et al, 2005):

> *For each feature in the input case, Find the corresponding feature in the stored case base Compare the two values to each other and compute the degree of match multiply by a coefficient representing the importance of the feature to the match Add the results to drive an average match score*

This number represents the degree of match of the old case to the input.

The nearest neighbor algorithm can be represented in the following equation (Watson & Marir, 1994).

$$\mathrm{NN}\ (\mathrm{I}, \mathrm{R}) = \frac{\sum_{i=0}^{n} wi \ \mathrm{x}\ \mathrm{sim}(f_i^{I}\ , f_i^{R})}{\sum_{i=0}^{n} wi}$$

Where: ***w*** is the importance weighing of an attribute, **I** is the target case, **R** is source case, **i** is individual attributes from 1 to n, *sim* is the local similarity function, and *f{* $^{\text{and}}$ h^{R} are the values for attribute *i* in the input case (I) and case in the case base (R) respectively, and ***n*** is the number of attributes in the case base.

Induction Retrieval

Depending on the size of the case base, the information amount contained in single cases, and the complexity of the used similarity measure, the retrieval step is often a challenging task with respect to computation time. In order to manage this complexity, a large number of different retrieval strategies have been developed (Schumacher & Bergmann, 2000). In inductive retrieval, use past cases to extract rules or construct decision (Lang & Lau, 2002). This technique finds target case-based on index source case. Cases are divided into a decision tree structure. Inductive retrieval used to retrieve set of matched cases and then nearest-neighbor retrieval rank these cases according to their similarity with target case.

2.2.2. Reuse

After selecting one or several similar cases, the reuse step tries to apply the contained solution information to solve the new problem. Often a direct reuse of a retrieved solution is impossible due to differences between the current and the old problem situation. Then the retrieved solutions have to be modified in order to fit the new situation. How this adaptation is performed strongly depends on the particular application scenario (Wilke & Bergmann, 1998).

In general, adaptation methods require additional general knowledge about the application domain. Because this leads to additional knowledge acquisition effort, many CBR systems used today do not perform case adaptation automatically, but leave this task to the user. Then, of course, the quality of the retrieval step influences the problem-solving capabilities of the entire CBR system primarily. Even if automatic adaptation is provided, the qualities of the retrieval result will strongly influence the efficiency of the system due to its impact on the required adaptation effort.

After adapting the retrieved case automatically or manually to fit the current situation, a solved case is obtained containing a suggested solution for the current problem.

2.2.3. Revise

Depending on the employed adaptation procedure, the correctness of the suggested solution often cannot be guaranteed immediately. Then it becomes necessary to revise the solved case. How such a revision is performed, strongly depends on the particular application scenario. For example, it might be possible to apply the suggested solution in the real-world to see whether it works or not. However, often a direct application of an uncertain solution is impossible due to the corresponding risks. Then the revision has to be performed manually by a human domain expert or by alternative methods such as computer simulation. Usually, the focus of the revise phase lays on the detection of errors or inconsistencies in the suggested solution and the initiation of further problem-solving attempts (Aamodt & Palaza, 1994).

2.2.4. Retain

If the solved case has passed the revise step successfully, a tested/repaired case will be available representing a new experience that might be used to solve similar problems in the future. The task of the CBR cycle's last step is to retain this new case knowledge for future usage. Therefore, the new case may be added to the case base. In most cases, a general storage of all generated cases is not always useful. In order to enable better control of the retain process, various approaches for selecting cases to be retained have been developed (Lenz *et al.*, 1998; Ferrario & Smyth, 2000). These approaches often imply a reorganization of the entire case base when adding a new case, for example, by removing other cases.

Generally, the capability to acquire new case knowledge during a CBR system's lifetime principally adds these systems to the class of learning systems. Conversely, many CBR systems developed so far do not exploit this concept of the CBR cycle at all. This holds true especially for the commercially employed systems. Further, the original idea of the CBR cycle focuses on learning case knowledge.

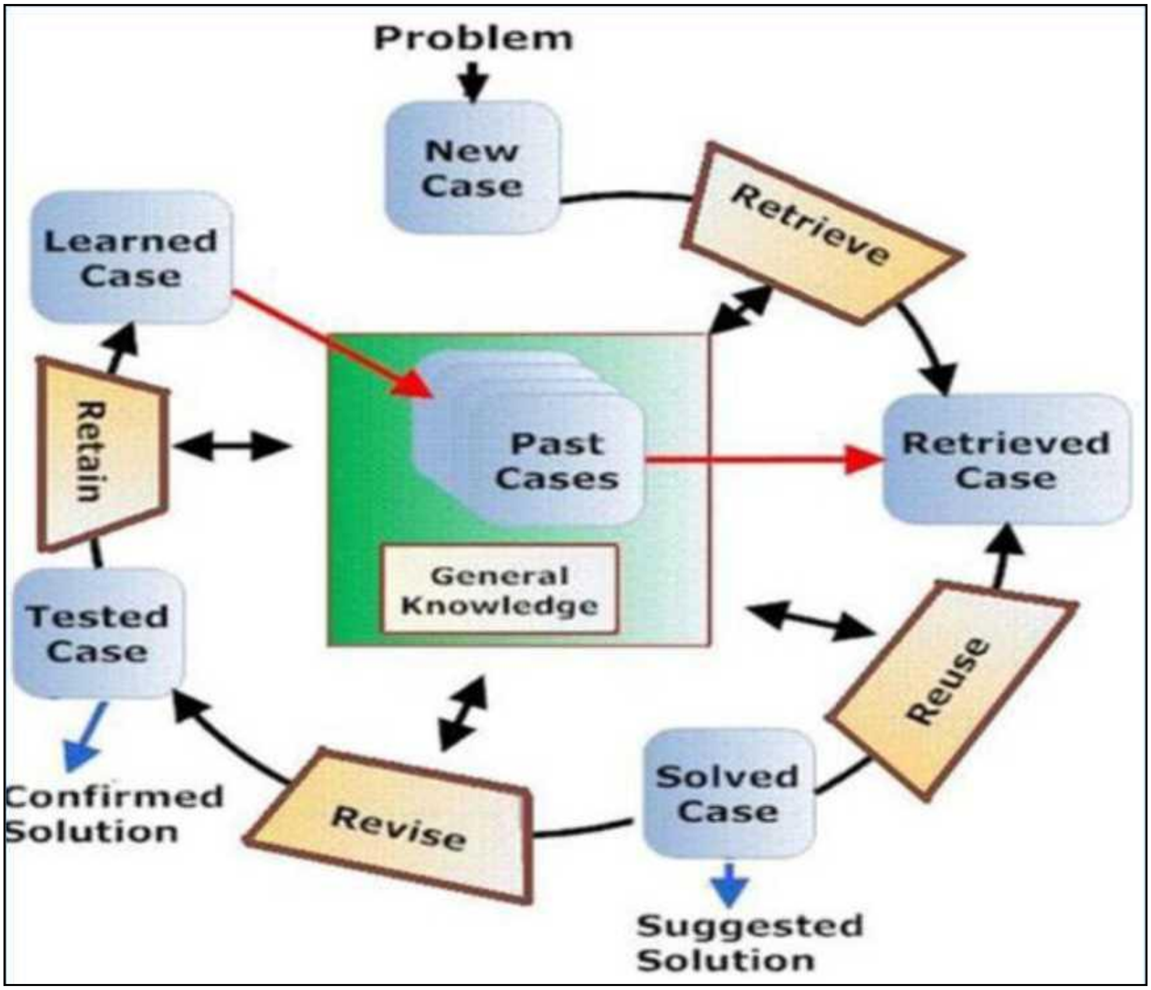


Figure 2. 1 CBR Life Cycle, Introduced by Aamodt & Plaza (1994)

2.3. Steps in CBR system design and development

To conduct a CBR project it is important to have a well-organized development procedure. According to Chan *et al*. (2000) there are four (4) steps for CBR system design and development.

Step 1: Domain Knowledge acquisition: in this step, a lot of effort is made in order to understand the problem domain. Information about the study domain and criteria's and procedures used are also collected in this step. A complete study should also include interviews with experts and consultants and a collection of some initial cases.

Step 2: Case representation: in this step, the software to be used for knowledge representation should be selected. The next step is to describe the case.

Step 3: System implementation: this describes the final system including the database of cases and the process of indexing and retrival within the chosen software.

Step 4: Verification and validation: in this step, some informal verification and validation should be conducted (Chan et al. 2000). Verification aims at demonstrating the consistency and correctness of the software (Adrion *et al*., 1982).

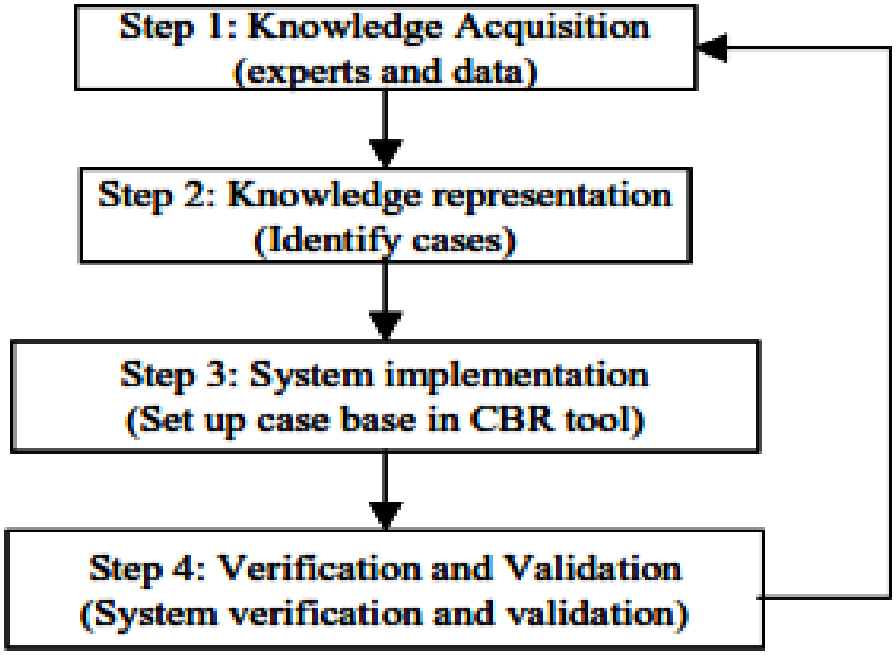

Step 1: Knowledge Acquisition
(experts and data)
Step 2: Knowledge representation
(Identify cases)
Step 3: System implementation
(Set up case base in CBR tool)
Step 4: Verification and Validation
(System verification and validation)

Figure 2. 2 Case based system development procedure (Adapted from Chan et al. 2000).

2.4. CBR System Performance Evaluation Methods

Evaluation of knowledge base system includes both system performance (statistical analysis) and user acceptance (Buchanan & Forsythe, 1991). The statistical analysis for CBR can be conducted for both retrieval and reuse process. The first task of CBR is to retrieve cases that are relevant to the new case (Aamodt & Plaza, 1994). As retrieval task of the CBR aims to retrieve cases relevant cases from the case base, precision and recall are useful measures of retrieval performance in CBR (McSherry, 2001). Recall is defined as the ratio of the number of relevant cases returned to the total number of relevant cases for the new case in case base (McSherry, 2001). Whereas precision is the ratio of the number of relevant cases returned to the total number of cases for a give new case (McSherry, 2001).

Only system performance evaluation based on statistical analysis does not assure the applicability of the system in the real life. Even though system that achieves better system performance statistically, it may not be comfortable to the user in solving the particular problem (Buchanan & Forsythe, 1991). As a result of this user acceptance is conducted to assess the applicability of the system for the real life.

2.5. CBR Tools

There are different types of tools that can be used for developing a CBR system. Most of these tools are commercial and few of them are non-commercial. The following CBR tools are indicated on the paper of Ashraf & Iqbal (2006) and Watson & Marir (1994).

ReCall

This CBR tool is written in C++ language. It provides both nearest neighbor and inductive retrieval algorithm. It can run on windows and UNIX workstations under Motif, Sun, HP series 700 and DEC Alpha, designed in open architecture that allows the user to add CBR functionality in the application.

Remind

It is produced by Cognitive Systems Inc. It is basically developed for Macintosh, but after some time, it is also developed for Windows and UNIX. ReMind offer template, nearest neighbor, inductive and knowledge-based retrieval. Its limitation is retrieving speed. Nearest neighbor is very slow, on the other hand inductive retrieval is very fast. When it creates inductive index, then it becomes slow (Watson & Marir, 1994). It will able to access data in ODBC-compliant databases and very influential tool.

ART*Enterprise

It is the product of Inference Corporation's (Watson & Marir, 1994). In 1980, ART was advertised as an AI-based tool. Inference dropped the label of AI and now ART*Enterprise marker as an integrated and object-oriented development tool.

ART* Enterprise has many features (Watson & Marir, 1994). It offers cross-platform support for all operating system. It provide excellent environment to integrate CBR with others application. Since the whole package is very powerful, ART*Enterprise is ideal tool for embedded CBR functionality within a corporate wide information system. Its package include graphic user interface (GUI) builder.

CasePower

Inductive Solutions Inc. developed CasePower tool. That tool builds its cases in matrix environment provided by Microsoft Excel. Rows and columns of spread sheet are used to define cases and their attributes. It uses nearest neighbor retrieval and reduces the search time by calculating the index in advance. If new case is retained, then entire set of case indices must be recalculated (Watson & Marir, 1994).

CBR-Express

This CBR tool primarily designed for help desk applications by Inference Corporation. It provides a comfortable user interface and fast retrieval speed. It has simple case structure and nearest neighbor retrieval algorithm of cases. The key features of CBR-Express are to handle free-form text and enable customers to describe their problem in own words. The use of trigrams means that CBR-Express is reluctant of spelling mistakes and typing errors such as letter transpositions. It stored cases in relational database. CBR-Express is network ready and cases can be shared between organization's networks.

Kate

This tool is developed by Watson & Marir (1994) that can run on MS Windows, Mac, or SUN. Kate is made up of Kate-induction, Kate-CBR, Kate-Editor and Kate-Runtime, this tool support both kind of Nearest Neighbor and Inductive retrieval algorithm. Kate-Induction is an ID3-based induction system that supports object-oriented representation of cases. Cases can be imported from many databases and spread sheets. Induction algorithm can make use of background knowledge. In induction algorithm, retrieval using trees is extremely fast. Kate- CBR uses nearest-neighbor approach. It supports same case objects hierarchies as Kate- Induction. Two techniques can be combined in a single application. Users can customize similarity assessments.

CaseAdvisor

It is marketed by Sentenia Software at Frazier University in Canada (Watson & Marir, 1994). It is also developed by Inference's CBR product. This software has three parts (Watson & Marir, 1994). These are CaseAdvisor Authoring, CaseAdvisor Resolution and CaseAdvisor WebServer

Eclipse - The Easy Reasoner

It is a product of Haley Enterprises. Eclipse is implemented in C by NASA. In late 1980, the former chief scientist of Inference developed a new language like Eclipse. Eclipse is available for Dos, Windows and UNIX platforms. It supports nearest neighbor and inductive retrieval (Watson & Marir, 1994).

Esteem

It is from Esteem Software (Watson & Marir, 1994). It is developed in Intellicorp's Kappa- PC. Esteem uses kappa's inference engine for developers to create adaptation rules. It supports application which has multiple case-bases and nested cases. This means that one can reference another case-base through an attribute slot in a case.

Casuel

European INRECA project developed Casuel (Watson & Marir, 1994). It is a common case representation language. Basically, it is interface between all INRECA component systems. Casuel is a flexible, object-oriented and frame like language.

Caspian

It is a CBR tool developed at the University of Aberystwyth in Wales (Watson & Marir, 1994). It can run on Ms-DOS and Macintosh. It has simple command line interface which can be integrated with a GUI front end if required. It performs simple nearest-neighbor match and uses rules for case adaptation.

JCOLIBRI

JCOLIBRI is a technological evolution of COLIBRI and it is an object-oriented framework in Java which is designed for building CBR systems. It is a java-based and uses JavaBeans technology for case representation and automatically generation of user interface. This framework is developed by the GAIA artificial intelligence group in Complutense University in Madrid. The framework is built in two hierarchical levels- upper and lower. The lower level consists of library of classes (Software modules) for full 4REs CBR cycle, also for definition of cases, attributes and connectors for access to outer databases. The upper level is “black box” - graphical interface, which allows non-complicated user CBR application generation based on lower level’s modules.

JCOLIBRI supports full CBR cycle. At the retrieve stage the nearest N cases are retrieved. At reuse stage several methods for adaptation are available (direct proportion and also in ontology). At revise stage methods for revision of cases are realized, as well methods for new indices generation and methods for decision making (preference elicitation). At retain stage there are methods for query retaining to the case base for future use. JCOLIBRI allows retrieval from clustered and indexed case bases and submits program interfaces (connectors) to access text and XML files, as well standard and descriptive logic databases. These interfaces can be used for diagnostic systems database access. There are lots of CBR applications, developed on JCOLIBRI based: additional shells (abstract levels) for distributed CBR systems, statistical CBR systems, multi-agent supervisor systems for text file classification, and lots of CBR recommender systems.

2.6. Data Mining and knowledge discovery concepts

Data Mining and Knowledge Discovery is one of recent developments in line with data management technologies. It combines the fields of statistics, machine learning, database management, information science and visualization. It is an emerging field. Despite this, it is increasingly being used in the industry as a tool to study their customers and make smart decisions (Li & Lia o, 2011). Knowledge discovery from databases is defined as the process of identifying valid, novel, potentially useful and ultimately understandable patterns of data. One of the crucial steps in Knowledge discovery is Data Mining and often they are used as synonyms (Deshpande & Thakare, 2010). Data Mining is the process of discovering valuable information from large data stores to answer critical business questions. It unveils implicit relationships, trends, patterns, exceptions and anomalies that were hidden to human analysis. In today's highly competitive market environment customers are spoilt by choices. Banks need to be proactive in analyzing customer preferences and profiles and tune their products and services accordingly to retain customer base (Bhambri, 2011). By segmenting customers into bad customers and good customers, bank can cut losses before it is too late (Kazi & Ahmed, 2012). By analyzing patterns of transactions, bank can track fraud transactions before it affects its profitability (Li & Lia o, 011). These are highly desirable capabilities where data mining could help.

Data mining is the process of deriving knowledge hidden from large volumes of raw data. The knowledge must be new, not obvious, must be relevant and can be applied in the domain where this knowledge has been obtained.

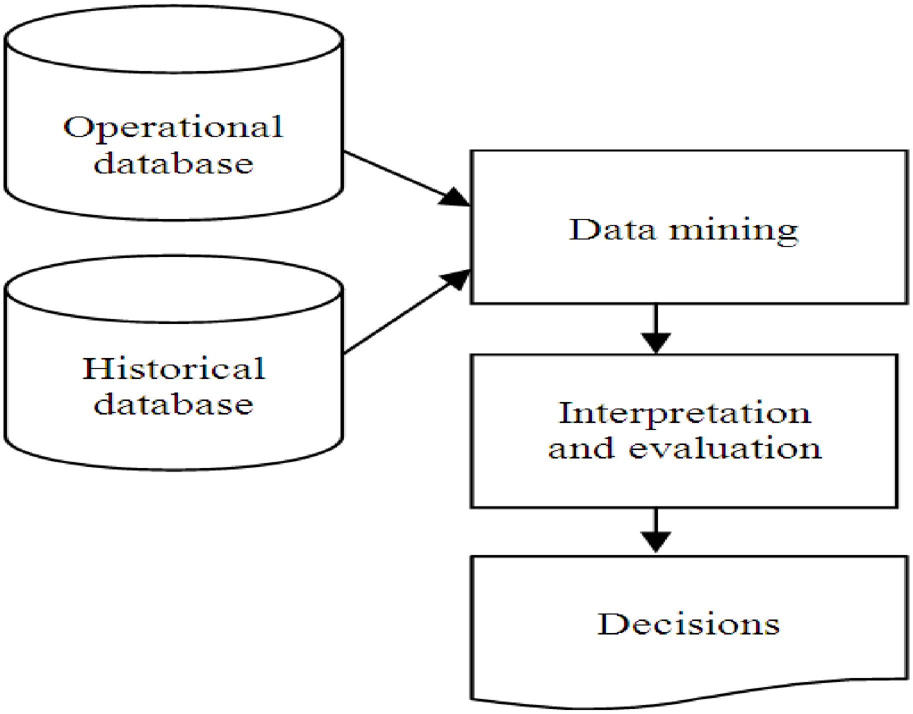


Figure 2. 3 Decision making with data mining (Pulakkazhy & Balan, 2013).

2.6.1. Data mining process models

There are different DM process model standards. The six step Cios *et al.* (2000) model, KDD process (Knowledge Discovery in Databases) and CRISP-DM (Cross Industry Standard Process for Data Mining) are some of the models that are used in different DM projects.

2.6.1.1 The six step Cios et al. model

This was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps (Cios & Kurgan 2005).

Understanding of the problem domain: In this step one works closely with domain experts to define the problem and determine the research goals, identify key people, and learn about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the DM goals, and include initial selection of the DM tools.

Understanding of the data: This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DM goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

Preparation of the data: This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data will be used as input for DM tools of step 4, is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be, further processed by feature selection and extraction algorithms (to reduce dimensionality), and by derivation of new attributes (using discretization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

Data mining: This is another key step in the knowledge discovery process. Although it is the DM tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned DM tools and selection of the new ones. DM tools include many types of algorithms, such as neural networks, clustering, preprocessing techniques, Bayesian methods, machine learning, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.

Evaluation of the discovered knowledge: This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire DM process may be revisited to identify which alternative actions could have been taken to improve the results.

Using the discovered knowledge: This step is entirely in the hands of the database owner. It consists of planning where & how the discovered knowledge will be used. The application area in the current domain should be extended to other domain.

2.6.1.2 The KDD process model

KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database as presented by Azevedo & Santos (2008). It is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions. Generally, there are five steps in the KDD process as discussed in chapter one namely data selection, data processing, data transformation, data mining and interpretation or evaluation (Azevedo & Santos 2008).

As indicated above, a KDD process involves preprocessing data, choosing a data-mining algorithm, and post processing the mining results. There are very many choices for each of these stages, and non-trivial interactions between them. Therefore both novices and DM specialists need assistance in KDD processes.

2.6.1.3 The CRISP-DM process

CRISP-DM (Cross Industry Standard Process for Data Mining), process model was first established by four companies in the late 1990s (Chapman *et al.*, 2000; Kurgan & Musilek, 2006; Azevedo & Santos, 2008).

According to CRISP-DM (Cross Industry Standard Process for Data Mining) the life cycle of a data-mining project consists of six phases (Chapman *et al.*, 2000; Azevedo & Santos, 2008). The sequence of the phases in the CRISP-DM process is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The CRISP-DM process has six stages.

Business understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

Data understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data preparation: It covers all activities to construct the final dataset from the initial raw data.

Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the DM results should be reached.

Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.6.2. Data mining Classification Techniques and Algorithms

Classification is the most important and popularly used technique in data mining. It is a process of finding a set of models or pre-defined conditions that describe and distinguish data classes or concepts. Classification is the derivation of a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database (Yongjian, 2006). As mentioned by Thair Nu Phyu (2009), classification is also called supervised learning. It is called supervised learning because it works on labeled attributes in which there is a specially designated attribute and the aim is to use the data given to predict the values of that attribute for instances that have not yet been see (Yongjian, 2006),

The given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label (classify) a new coming pattern. Classification technique maps data into predefined groups. Classification is a two step process consisting of model construction and model usage. In the first step, a classifier is built describing a predetermined or labeled set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or learning from a training set made up of database instances and their associated class labels. This step is called model construction (Deshpande & Thakare, 2010). Generally, classification is a process of building model that describe data class and used to predict the class of objects whose class label is unknown. It finds out the relationship between predictor value and the target value. The model is based on the analysis of a set of training data. The data; historical, for a classification is typically divided into two datasets: one for building the model; the other for testing the model. Thus the various classification approaches can be employed on credit data for obtaining specific information. Decision tree, K-nearest neighbor, Byes classifier, neural network, support vector machine and rule based learning are some of the classification data mining techniques. In this report, decision tree, Byes classifier and rule based learning (production rules) are discussed.

I. Decision Tree

A decision tree is predictive modeling technique used in classification, clustering, and prediction tasks. It uses a divide and conquers technique to split the problem search space into subsets" (Dunham, 2000). A decision tree is a classifier expressed as a recursive partition of the instance space. Decision tree is used in data mining to classify objects into values of the dependent variable based on the values of independent variables. According to Fekadu (2004), there are two main types of decision trees. These are classification trees and regression trees. Classification trees are decision trees used to predict categorical variables, because they place instances in categories or classes. And, the second one is regression trees, which is a decision tree used to predict continues variables (variable which are not nominal). Classification trees can provide the confidence to correctly classify the data. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class. On the other hand, regression trees estimate the value of a target variable that takes on numeric value. The structure of decision tree is a tree like structure, where each internal node represents a test on an attribute, each branch characterizes an outcome of the test, and leaf nodes at the end represent classes in which the data is assigned. The top most nodes in a tree are the root node.

The basic algorithm for decision tree induction is greedy algorithm that constructs decision trees in a top-down recursive divide-and –conquer manner (Thair Nu Phyu , ,2009). The algorithm is summarized as follows.

```
Create a node N;
If samples are all of the same class, C then
    Return N as a leaf node labeled with the
class C; If attribute-list is empty then
    Return N as a leaf node labeled with the most common class in samples;
    select test-attribute, the attribute among attribute-list with the highest
    information gain;
    label node N with test-attribute;
for each known value $a_i$ of test-attribute
    grow a branch from node N for the
    condition test-attribute= $a_i$;
    let $s_i$ be the set of samples for which test-attribute= $a_i$;
    If $s_i$ is empty then
        attach a leaf labeled with the most common class
    in samples; else attach the node returned by
    Generate_decision_tree(si,attribute-list_test-attribute)
```

ID3, C4.5 (J48) and CART, in their respective order of invention and usage, are algorithms used in decision tree construction. They adopt a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts a training set. There are different decision tree algorithms like Id3, J48graft, AD tree, J48 etc.

J48

J48 is an implementation of Quilan algorithm (C4.5). J48 classifier build a decision tree for the given data set, whose nodes represent discrimination rules acting on selective features by recursive partitioning of data using depth-first strategy.

The algorithm used each attribute of the data to make decision by splitting the data into smaller subsets. All the possible tests are considered during decision making based on information gain value of each attribute.

J48 algorithm is WEKA’s improved implementation of C4.5 algorithm. The process of the J48 algorithm to build a decision tree is as follows:

1. Choose an attribute that best differentiates the output attribute values.

2. Create a separate tree branch for each value of the chosen attribute.

3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.

4. For each subgroup, terminate the attribute selection process if:

    a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.

b. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a) label the branch with the output value seen by the majority of remaining instances.

5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

II. Rule based classification

Rule based classifiers group instances by using a set of IF…..THEN rules. Rules are comprised of Left Hand Side (LHS) also called antecedent or condition and Right Hand Side (RHS) also called rule consequent or conclusion (Datta & Saha, 2009). A given rule r covers an instance z if the attributes of the instance satisfy the condition (LHS) of the rule. Rule based classification techniques are divided into two namely; direct method and indirect method. Rules based classifiers which extract rules directly from data, for example RIPPER, are called direct methods (Datta & Saha , 2009).

Indirect methods are those that extract rules from other classification model like decision tree, for example C4.5 rules (Ali & Tickle, 2009). Direct methods first grow a single rule (Rule growing) then remove instances from this rule (Instance Elimination) after that prune the rule (Stopping Criterion and Rule Pruning) and then finally add rules to current rule set. PART and JRIP are algorithms which are rule based classifiers.

PART

PART is rule based classifier which generates rules repeatedly producing partial decision trees (Ali & Tickle, 2009).As cited by Datta & Saha (2009) , Frank & Witten (1998) stated that, the PART technique avoids global optimization in which pruning is effected after all rules are generated used in C4.5 and RIPPER. It builds a partial decision tree to obtain a rule using C4.5‘s procedures to build a tree. It identifies the rule that identifies many instances using separate and conquer then separate them out, repeat and makes the best leaf into a rule(Ali & Tickle, 2009), (Datta & Saha , 2009).

III. Bayesian Network Classifiers

Bayesian networks are graphical models which are very useful for representing variables (as nodes of the graph) and the probabilistic relationships between them (as connections, or edges of the graph). By knowing the value at one of the nodes in a Bayesian network, one can infer the value of other nodes in the network. Bayesian network classifiers are used in many fields and one common class of classifiers are NaiveBayes classifiers. The induction of classifiers from data sets of pre-classified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. One of the most effective Bayesian network classifiers, in the sense that its predictive performance is competitive classifiers, is the Naive Bayesian classifier (Choge, 2012). This classifier learns from training data the conditional probability of each attribute *Ai* given the class label *C*. Classification is then done by applying Bayes rule to compute the probability of *C* given the particular instance of *A*1,…., *An*, and then predicting the class with the highest posterior probability.

According to Bhambri (2011), on many real-world datasets naive Bayesian learning gives better test set accuracy than any other known method, including back propagation and C4.5 decision trees. Also, these classifiers can be learned very efficiently. Bayesian networks can have different advantages. Among those, some of them are provide probabilistic output, can operate with limited sensor data availability, more flexible relative to engineering development then traditional expert systems, used for both data qualification (state recognition) and anomaly reasoning, can operate in a central or distributed run-time environment either shore-side or ship-board. The reason why use bayesian networks is Bayesian inference methods have proven to be valuable for knowledge-based data mining applications, and are based on a causal (explanation based) modeling framework. Because relationships between variables in a Bayesian network are defined probabilistically, trends can be detected and analyzed over a continuous scale, rather than in a Boolean fashion.

2.6.3. Application areas of Data Mining in banking

Banking information systems contains huge volumes of data both operational and historical. Data mining can assist critical decision making processes in a bank (Ionita & Ionita, 2011). Banks who apply data mining techniques in their decision making hugely benefit and hold an edge over others who don’t. Some of these decisions are in the areas of marketing, risk management and default detection, fraud detection, customer relationship management and money laundering detection (Khac & Kechadi, 2010;Dheepa & Dhanapal,2009).These applications are described below.

- **Risk Management and Default Detection**

Every lending decision a bank takes involve a certain amount of risk. Knowing customers’ ability to repay can greatly enhance a credit manager’s decisions. Data mining can also help to identify which customer is going to delay or default a loan repayment (Kazi & Ahmed, 2012). This advanced knowledge can help the bank to take corrective measures to prevent losses. For such forecasting, parameters to consider are turnover trends, balance sheet figures, limit utilization, behavioral patterns and cheque return patterns. Data mining can derive this score using the past behaviors of the borrower related to debt repayments by analyzing available credit history (Chopra *et al*., 2011).

- **Marketing**

Marketing is one of the mostly used application area for Data Mining by the industry in general (Khac & Kechadi, 2010). Banking is not an exception. Retaining customers and finding new customers are getting increasingly difficult because of cut throat competition prevailing in the market these days. This is where data mining can help a great deal (Chopra *et al*., 2011). Data mining applied to customer relationship management systems can analyze customer data and can discover key indicators to help the bank to be equipped with the knowledge of factors that affected customer’s demands in the past and their needs in the future (Khac & Kechadi, 2010). This enables the bank to targeted marketing. Sequential patterns can be analyzed to investigate changing customer preferences and can approach customers pro-actively.

- **Fraud Detection**

Banks lose millions of dollars annually to various frauds. Detecting fraudulent transactions can help the banks to act early and limit damages. Fraud detection is the process of identifying fraudulent transactions from genuine transactions or in other words this process segregates a list of transactions into two classes namely fraudulent and legitimate (Khac & Kechadi, 2010). Most important area where fraud detection can help is the credit card products. Clustering methods can be used to classify transactions and outliers can be analyzed for frauds (Dheepa & Dhanapal, 2009).

- **Money Laundering Detection**

Money Laundering is the process of hiding the illegal origin of “black” money so as to legitimize it (Khac & Kechadi, 2010). Banks are commonly used as channels to launder money. Therefore governments and financial regulators require banks to implement processes, systems and procedures to detect and prevent money laundering transactions. Failure to detect and prevent such illegal transactions can invite hefty fines both monetarily and operationally which can prove very costly for the bank and even can make its survival difficult. Conventional rule-based transaction analysis based on reports and tools will not be sufficient to detect more complicated transaction patterns like surfing and networked transactions (Khac & Kechadi, 2010). Here data mining techniques can be applied to dig out transaction patterns.

2.6.4. Data mining tool selection

Data mining tools need to be versatile, scalable, capable of accurately predicting responses between actions and results, and capable of automatic implementation (Chackrabarti, *et al*, 2009). Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to the business strategies, knowledge bases, scientific and medical research (Han & Kamber, 2006).

Data mining tools predict future trends and behaviors and help organizations to make practical knowledge-driven decisions (Larose, 2005). The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were more time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations (Deshpande & Thakare, 2010).

There are factors that contribute to the usefulness of data mining tools or software to the intended data mining tasks: The tool selected should be able to provide the required data mining functions. The data mining functionality that the researcher has intended to carry out in this research is prediction. In addition the methodologies used by the data mining software to perform each of the data mining functions are also important factor to consider. The researcher has chosen the J48 decision tree classifier implementation that implements C4.5 algorithm.

Different researchers develop a framework to evaluate and select an appropriate data mining tools. However, the evaluation and selection of an appropriate data mining tool for this research was done based on certain criteria. The researcher had to first set criteria for tool selection

As a result the research used a frame work which is proposed by Collier et al. (1999) which consisting of different categories of criteria for evaluating and selecting data mining tools or software. The criteria used in this research to select one tool from the other were the following:

- Platform Variety (Does the software run on a wide-variety of computer platforms?)
- Performance of the tool in terms of speed and quality
- Algorithmic Variety (inclusion of various clustering and classification algorithms)
- The data mining tasks that the tool is intended for
- The compatibility of the tool to the operating system at hand (MS window)
- The possible formats for the data that is to be analyzed.

As a result, in this research Weka 3.6.5 software is used as a mining tool. In addition Microsoft Excel for data cleaning and for converting the original file to CSV file format, and Microsoft Word for documentation purpose have been used.

Weka includes varieties of tools, for preprocessing a data set, such as attribute selection, attribute filtering and attribute transformation (Witten & Frank, 2000).

In addition, the selected tool can comfortably operate on windows operating system and standalone environment. Hence Windows 7 operating system on a standalone machine has been utilized. Another important consideration in tool selection is visualization capabilities. The variety, quality and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of data mining systems. Weka has a facility to visualize its output in this regard (Witten & Frank, 2000).

It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code. The data provides any meaningful information that can be used to know anything about any object (Ionita & Ionita, 2011).

As noted by Witten & Frank, (2000), advantages of WEKA include the following:

- Free availability under the GNU General Public License
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
- A comprehensive collection of data pre-processing and modeling techniques
- Ease of use due to its graphical user interfaces

2.7. Related Works

A study conducted by Maria & Lusi (2002) proposed the use of case based reasoning (CBR) approach to support the prediction of bank lending decisions. The approach is applied to a survey of bank lending developed by euro system and conducted by the national central bank in each country. The main objective is to enhance the knowledge of credit standards and credit conditions in the euro area. The system uses the data from euro system survey for Portugal. This work describes the heuristic between past and new cases. A pilot study reveals that the system is capable of generating forecasts with a high level of accuracy. The results obtained lead to the conclusion that the system can forecast with considerable precision (90%) the decision of economic agents.

The experiments with hypotheses also show that the economic sentiment play a decisive role in both considered segments (enterprises and households) regarding the criterion for loan approvals and the other conditions underlying credit granting. The current research is different from this work with its attempt to use automated knowledge acquisition in addition to the manual knowledge acquisition mechanism and it considers the use of local data which uses a more or less different criterion for loan approvals and the other conditions underlying credit granting.

A study was conducted by Samuel, *et al*. (2012) to determine some risk factors that influence loan default repayment among customers in Akuapem rural bank. They used secondary data on some variables which influence customer loan default was obtained from the credit department of Akuapem rural bank. Data was collected for the period 2006 to 2010. A logistic regression model was fitted to the data. It was found that among the variables that were used, Security and Type of Loan were significant to the study whereas Sex, Marital Status, Age, Educational Level, Town were not significant to the study. They conclude that there is a high risk of customers who use personal guarantee to default than those who use collateral as a security in accessing the loan. Taking transport loan as a reference group, the risks of a customer defaulting when given a personal loan is less than when given a transport loan, all other factors being equal. The data was collected from the credit department of the bank, since they keep records of all the bank's loan customers and other relevant information. As the researchers view, this study serves as a preparatory ground for further analysis into the subject matter.

Choge (2012), did a research with the objectives to examine whether naïve Bayes Classifier can be applied accurately to consumer credit evaluation or not. The results of the study have shown that naïve Bayes Classifier can be used to evaluate credit applications. The classification accuracy obtained indicates that the naïve Bayes Classifier has the ability to correctly classify credit applications as either "good" or "bad". Identifying "bad" credit applications at an early stage ensures that there is reduced loss of revenue to the credit lending institution. Hypotheses testing have been utilized to show that features in the data, for instance, net income, age, number of credit cards, the time with account, can affect the performance of the classification system. As a result, feature selection has proven to be vital in improving the performance accuracy of the classifier. A model has been designed and used to evaluate the credit data. This ensures efficiency in credit evaluation processes that is free of bias and ensuring that the results are obtained in a short period of time. By utilizing such technique as naïve Bayes Classifier in credit evaluation as developed in this study, it will reduce any bias or emotional intention that can distort the decision process thereby leading to reduced cost of credit processing and improved quality of customer service. Furthermore, the high accuracy of this model proves that naïve Bayes Classifier can be useful for classifying credit applications. The current work is different from this work since it attempts to use different DM algorithms for model building and got a better result with J48 Classifier and uses the generated knowledge with different CBR framework called JCOLIBERI.

Based on the literature review the researcher conducted, it is clear that different scholar's different approaches which attempts to solve diversified problems within different domain. It is also noted that some scholars uses combined approaches other than a single approach which aims getting a better results. Even if different scholars used DM and KBS for Credit management previously with different area, data and technique, to the researcher knowledge there is no local research attempts made to use DM and CBR for credit approval.

As a result the researcher attempted to use a combined use of DM and CBR for credit risk decision making using local dataset. The researcher used different classification algorithms for model building and to discover detection and prediction patterns and used the one which registers a better result. It also used test instances which registers best accuracy as knowledge base for CBR development.

In addition to this, this research is different from other previous researches since it uses data mining technique for knowledge acquisition and uses the results with case based reasoning system for better results of credit approval decision making since the patterns detected using DM from the bank data set will help the bank to understand previous credit risk cause and to forecast future events that can help in its decision- making processes (Pulakkazhy & Balan, 2013). Historical default patterns can also help in predicting future defaults when same patterns are discovered (Costa *et al.*, 2007). Due to this fact, the researcher used automatically generated knowledge with case based reasoning for credit approval decision making.

CHAPTER THREE

KNOWLEDGE ACQUISITION, MODELING AND KNOWLEDGE REPRESENTATION

The development of an efficient knowledge-based system (KBS) involves the development of an efficient knowledge base that has to be complete, coherent and non-redundant. The step of knowledge acquisition is one of the major bottlenecks in the stage of case base development. Usually, for each application domain there are several sources of knowledge (human experts, the specialized literature which includes textbooks, books, reviews, collection massive of data, etc) (Asghar & Iqbal, 2009).

According to Asghar & Iqbal (2009), in order to make knowledge extraction as much as correct as possible (i.e. in order to keep the correctness of the knowledge as it is kept at the source) different techniques could be applied. Among these techniques, data mining techniques and, more general, knowledge discovery techniques became the most used in the recent years. The researcher acquires knowledge using two types of knowledge acquisition methods which are manual and automatic knowledge extraction.

3.1. Manual knowledge acquisition

Knowledge can be acquired from domain experts (also called tacit knowledge) and from documents, which is codified knowledge (also called explicit knowledge). The researcher used different documents which are Guideline for credit approval process, evaluation and approval of credit in commercial bank of Ethiopia which is prepared by National Bank of Ethiopia (NBE) and Commercial bank of Ethiopia (CBE).

The researcher also conducted domain expert interview with purposively selected Loan Officers, Customer Relationship Mangers and Credit Analysts of CBE that are working in the Credit Section for domain know how. These experts selected purposefully for extensive discussion using structured and unstructured interviews to understand the domain knowledge and to verify the cases acquired from the previous customer credit history. These experts are practically participating throughout the research work, and they are consulted to confirm the correctness of the acquired knowledge. In addition, secondary source of knowledge has been gathered from the internet, banks credit guidelines, manuals, research papers and journal articles. But, the primary data source is credit cases of Commercial Bank of Ethiopia. The researcher founds that the knowledge acquired from documents and interview are similar and the researcher used the result as data triangulation. The results obtained from the interview and document analysis are presented below.

3.1.1 What is credit risk?

Credit risk is defined as the probability that some of a bank�once assets, especially its loans, will decline in value and possibly become worthless. Because banks hold little owners 'capital relative to the aggregate value of their assets, only a small percentage of total loans need to go bad to push a bank to the brink of failure. Credit risk is the risk of a loss resulting from the debtor's failure to meet its obligations to the Bank in full when due under the terms agreed (Mirach, 2010). Thus, management of credit risk is very important and central to the health of a bank and indeed the entire financial system. As banks make loans, they need to make provisions for loan losses in their books.

Credit Risk is the potential that a bank's borrower or counterparty will fail to meet its obligations in accordance with agreed terms. Thus credit risk arises from non-performance by borrower or a counter party due to either inability or unwillingness to perform as per the contracted.

According to the experts opinion, across country experience evident that credit activities are the main determining factors for the wellbeing of the financial sector's, especially in intermediation activities such as banking services. As discussed earlier, loans are the largest and most obvious source of credit risk and hence, ensuring prudent lending operation that reflects an acceptable risk reward ratio is, therefore, an area in which banks have to devote considerable skills and research. Thus, management of credit risk is very important and central to the health of a bank and indeed the entire financial system. As banks make loans, they need to make provisions for loan losses in their books. The higher this provision becomes, relative to the size of total loans, the riskier a bank becomes. An increase in the value of the provision for loan losses relative to total loans is an indication that the bank‘s assets are becoming more difficult to collect.

3.1.2. Types of Credit

Based on the banks credit guideline there are 3 basic types of credit. By understanding how each works, financial institutions will be able to get the most solution for their loan recovery and avoid paying unnecessary charges.

- **Service credit** is monthly payments for utilities such as telephone, gas, electricity, and water. You often have to pay a deposit, and you may pay a late charge if your payment is not on time.
- **Loans**:   Loans can be for small or large amounts and for short or long periods. Loans can be repaid in one lump sum or in several regular installment payments until the amount borrowed and the finance charges are paid in full. Moreover, loans can be secured or unsecured.

- **Installment credit:** is described as buying on time, financing through the store or the easy payment plan.

3.1.3. Credit Management

Credit risk has the highest weight among risks taken by the Bank in the course of its banking activities. Credit management is implementing and maintaining a set of policies and procedures to minimize the amount of capital tied up in debtors and to minimize the exposure of the business to bad debts.

Credit Management, from a debtor’s point of view, is managing finances especially debts so as not to have a tail of creditors lurking behind your back. Credit management is a responsibility that both the debtor and the creditor should seriously take. When it functions efficiently; credit management serves as an excellent instrument for the business to remain financially stable (Mirach, 2010).

The main tool to restrict and control the credit risk taken by the Bank is the credit limit system.

The following types of credit risk limits are put in place:

- counterparty limits;
- limits for independent risk-taking by the Bank's branches; and
- Credit risk limits by countries/industries/regions.

Credit risk limits are determined by the Credit Committee and approved by the Bank's Management Board (in case the Credit Committee does not have the required authority). A part of authorities for putting credit limits in place is delegated to Branch Credit Committees (for standard credit operations within the special limit for independent credit risk-taking by branches), as well as to the Small Credit Committee and the Moscow Region Credit.

3.1.3.1. Basic Requirements for Credit Decisions

According to the experts interview results , Commercial Bank of Ethiopia has uniform basic requirements which applicants are expected to present and become eligible for loan. Thus, credit decisions of the Bank are based on the fulfillment of these requirements as mentioned below.

- Renewed Trade license
- Marriage certificate for mortgagors or Confirmation letter from authorized office
- Loan application letter stating type, amount, purpose and term of repayment of loan requested and type of proposed collateral and others.

So as to qualify for credit, every applicant should fulfill the aforesaid requirements. If there is match between the documents provided by the client with that of the requirements set by the bank on the check list, the client will be eligible for the loan.

In the lending process, as per the interview conducted with the District manager, and loan officers the Bank prefers the business type and applicant creditworthiness as first way out and collateral is the second way out as basis for lending.

In principle, loan can be provided both on clean base and on collateral base. However, the Bank prefers collateral based lending because of the following main reasons

- The economic level of the country: the living standard of the society, poverty, etc;
- The culture of the society in lending is at its infant stage
- The educational level
- Limited resources of the bank this is to minimize the shortage of finance
- It is believed to be the safest way of lending in minimizing credit risk

3.1.4. Credit Analysis

Credit analysis is the primary method in reducing the credit risk on a loan request. This includes determining the financial strength of the borrowers, estimating the probability of default and reducing the risk of non-repayment to an acceptable level. In general, credit evaluations are based on the loan officer's subjective assessment (or judgmental assessment technique).

Once a customer requests a loan, bank officers analyze all available information to determine whether the loan meets the bank's risk-return objectives. Credit analysis is essentially default risk analysis, in which a loan officer attempts to evaluate a borrower's ability and willingness to repay.

A bank's credit analysts often use the five C's of credit to focus their analysis on the key dimensions of an applicant's credit worthiness.

Lawrence (1997) identified five C's of credit. They include; Character, Capacity, Capital, Collateral, and Conditions.

i. **Character:** The applicant's record of meeting past obligations, financial, contractual, and moral. Past payment history as well as any pending or resolved legal judgments against the applicant would be used to evaluate its character.
ii. **Capacity:** The applicant's ability to repay the requested credit. Financial statement analysis, with particular emphasis on liquidity and debt ratios, is typically used to assess the applicant's capacity.
iii. **Capital:** The financial strength of the applicant as reflected by its ownership position. Analysis of the applicant's debt relative to equity and its profitability ratios are frequently used to assess its capital.

iv. **Collateral**: The amount of assets the applicant has available for use in securing the credit. The larger the amount of available assets, the greater the chance that a firm recover its funds if the applicant defaults. A review of the applicant's balance sheet, asset value appraisals, and any legal claims filed against the applicant's assets can be used to evaluate its collateral.

v. **Conditions:** The current economic and business climate as well as any unique circumstances affecting either party to the credit transaction. For example, if the firm has excess inventory of the items the applicant wishes to purchase on credit, the firm may be willing to sell on more favorable terms or to less creditworthy applicants. Analysis of the general economic and business conditions, as well as special circumstances that may affect the applicant or firm is performed to assess conditions.

The credit analyst typically gives primary attention to the first two C's-character and Capacity- because they represent the most basic requirements for extending credit to an applicant. Consideration of the last three C's-Capital, Collateral, and Conditions- is important in structuring the credit management and making the final credit decision, which is affected by the credit analyst's experience and judgment.

3.1.5. Assessment of Applicants Creditworthiness

Once a customer requests a loan, bank officers analyze all available information to determine whether the loan meets the bank's risk-return objectives. Credit analysis is essentially default risk analysis, in which a loan officer attempts to evaluate a borrower's ability and willingness to repay. The Bank assesses the creditworthiness of a loan applicant mostly by gathering detail information with regard to:

A. The applicant

- Whether the applicant is customer of any other bank. This is done to check whether the applicant has any loan arrear with other banks. This will be checked by the help of NBE Central Database of the credit information center.
- The exposure of the applicant to credit and his track record in meeting his obligation.
- The educational level and experience of the applicant.
- The character, capacity of the applicant and his social acceptance in trustworthiness.

B. Collateral

- Credit policy of the Bank, for building collateral (85%), For Vehicles and Machinery (70%), for cash (100%), Merchandise (70%), Leased land (30% less lease amount if construction not started), Treasury bills and Government bonds,(100%), etc.
- Marketability and habitability
- Easily transferability

C. Business viability

- Based on the basic financial measurements used to certify the credit worthiness of the business the Bank depends on liquidity rate, solvency, efficiency ratio, sales turnover and profit margin of the business. Once the Bank assessed the creditworthiness of the applicant, the credit decision flow is as depicted below.

3.1.6. Provisions

Loans and advances are financial instruments originated by the bank by providing money to the debtors. It is stated at costless impairment losses. Impairment losses comprise specific provisions against debts identified as bad and doubtful and general provisions against losses which are likely to be present in any loans and advances portfolio. The Bank follows the National Bank of Ethiopia Supervision of Banking Business Directives SBB/43/2008 in determining the extent of provisions for impairment losses. The Directive classifies loans and advances into the following.

a) Current or Pass Loans

Loans and advances in this category are fully protected by the current financial and paying capacity of the borrower and are not subject to criticism. It is fully secured, both as to principal or interest payments, by cash or cash substitutes are classified under this category regardless of past due or other adverse credit factors.

b) Special Mention

Any loan or advance past due 30 days or more, but less than 90 days.

c) Substandard

Non- performing loans or advances past due 90 days or more but less than 180 days.

d) Doubtful

Non- performing loans past due 180 days or more but less than 360 days.

e) Loss

Non-performing loans or advances past due 360 days is classified as loss.

3.1.7. Credit assessment and Risk grading

- **Credit assessment**

A thorough credit and risk assessment should be conducted prior to the granting of loans, and at least annually thereafter for all facilities. The results of this assessment should be presented in a Credit Application that originates from the relationship manager/account officer (RM), and is approved by Credit Risk Management (CRM). The RM should be the owner of the customer relationship, and must be held responsible to ensure the accuracy of the entire credit application submitted for approval. RMs must be familiar with the bank‘s Lending Guidelines and should conduct due diligence on new borrowers, principals, and guarantors.

- **Risk grading**

All Banks should adopt a credit risk grading system. The system should define the risk profile of borrower‘s to ensure that account management, structure and pricing are commensurate with the risk involved. Risk grading is a key measurement of a Bank‘s asset quality, and as such, it is essential that grading is a robust process. All facilities should be assigned a risk grade. Where deterioration in risk is noted, the Risk Grade assigned to a borrower and its facilities should be immediately changed. Borrower Risk Grades should be clearly stated on Credit Applications.

The more conservative risk grade (higher) should be applied if there is a difference between the personal judgment and the Risk Grade Scorecard results. It is recognized that the banks may have more or less Risk grades; however, monitoring standards and account management must be appropriate given the assigned Risk Grade.

3.1.8. The Commercial Credit Approval Process

In its simplest form, a Bank makes money by taking funds in the form of low cost deposits (checking/savings/money market accounts) and then loaning out that same money at a higher interest rate. A bank’s profit is a result of the “spread” or the difference between the rate it pays for the deposits and the return that it makes on loans. In order to protect the deposits entrusted to the bank, all loans go through a credit approval process. During this process a bank determines whether or not a proposed loan has a high enough chance of being repaid. If a commercial loan request doesn’t make the cut then it is denied or restructured to reduce the risk to the bank.

A thorough credit and risk assessment should be conducted prior to the granting of loans, and at least annually thereafter for all facilities. The results of this assessment should be presented in a Credit Application that originates from the relationship manager/account officer (RM‖), and is approved by Credit Risk Management (CRM). The RM should be the owner of the customer relationship, and must be held responsible to ensure the accuracy of the entire credit application submitted for approval. RMs must be familiar with the bank‘s Lending Guidelines and should conduct due diligence on new borrowers, principals, and guarantors. It is essential that RMs know their customers and conduct due diligence on new borrowers, principals, and guarantors to ensure such parties are in fact who they represent themselves to be. All banks should have established Know Your Customer (KYC) and Money Laundering guidelines which should be adhered to at all times.

3.1.8. Key Players in the Credit Approval Process

When a potential borrower makes a request for a loan (in this case, we are going to assume that this is a commercial loan), there are many individuals involved in the decision to approve or deny the loan. Let’s take a look at the key players.

Relationship Manager: This is the “sales” person. Their primary responsibility is to bring new “relationships” to the bank, which includes both deposit and loan accounts. These are the people who are out actively calling on accounts, taking prospects to lunch, playing golf, and generally doing anything necessary to get the customer to move their banking relationship to whatever bank they represent. Most of the time, the loan request is brought into the bank through the relationship manager.

Credit Officer: The credit officer is the individual with the authority to approve or deny the loan request. Each credit officer has a certain amount of “approval authority” indicating the maximum Birr amount that they are able to approve. Depending on the size of the loan request, the local credit officer may be able to approve it. If not, it may need to go to a more senior credit officer or board of directors if needed.

Credit Analyst: This is the number cruncher. This individual analyzes all of the information gathered by the relationship manager and puts it in a loan approval document. This document outlines all of the risks and benefits of approving the loan and outlines the feasibility of repayment. When a loan is approved, this document is signed by both the relationship manager and the credit officer. All of these individuals work together towards an approval decision on the loan.

3.1.8.1. How Loans are made?

The process begins when a relationship manager identifies a loan prospect and begins preliminary discussions with the customer about the loan or the customer him/her applies for a loan. These discussions typically include the amount, term, and rate of the loan.

Next, the relationship manager presents the request to their sales manager (at branch level, there might not be a sales manager, at that time branch manager take the responsibility). They have a discussion about whether or not this is a customer that they would like to have. The relationship manager will leave this discussion with a directive as to proceed with the request or not.

If they get the go-ahead from the sales manager, they will go back to the customer and request several documents to use in evaluating the request. At a minimum, this usually includes: 2 years of tax returns for both the business and the individual, a personal financial statement, a current financial statement for the business. Marriage certificate, tax clearance, Collateral document and any other documents that will support the loan request.

The relationship manager will take all of the above documentation and submit it to the Central credit processing center. The central processing team then sends the request for 2 different teams for detail analysis namely Appraisal team and Relationship team.

The credit analyst will then begin their analysis, which takes the form of a loan approval document. Let's take a look at the key components of a loan approval document, sometimes called a "Credit Memo" internally.

i. Summary of the Loan Request: The summary section of the credit memo provides a high level overview of the request, and will include the loan amount, loan term, proposed interest rate, individual/corporate guarantors, etc.

ii. Information About the Customer / Company : This section includes information on who the principals are, how much operating experience they have, how long the company has been in business, what the company does, etc.

iii. Repayment Analysis: This is the most important section. In this section, the credit analyst will take all of the financial data that they have received and make an attempt to estimate the Customer's ability to repay the loan. This includes analyzing past data, looking at growth trends, industry trends, proposed loan terms, and certain assumptions that will get them to a number called the DSCR or Debt Service Coverage Ratio. The DSCR is the company's free cash flow divided by the estimated debt service on the proposed loan.

iv. Company/Borrower Financial Analysis: This section will explore the financial statements of the proposed borrower. It will look at historical trends, critical ratios, and interim data to determine the financial health of the borrower.

v. Individual/Guarantor Financial Analysis: Many bank loans require the individual guarantee of the company principal(s). As such, an analysis is performed on each of the individuals who will be guaranteeing the loan; analyzing their ability to cover any shortfalls in the debt service should things not go as planned. The key here is guarantor liquidity or how much cash they have in the bank, and excess personal cash flow.

vi. Relationship Analysis: This analyzes the customer's relationship with the bank. Are they a current customer? Do they have deposits? What other loans are outstanding?

Once the credit analyst has completed the loan approval document, they will send to central approval team by including their results and the team recommendations. Typically there will be 2 or 3 rounds of edits including the proposed loan covenants in the document.

The same procedure is carried out with Relationship team members including analyzing the customer credit document, Collateral audit and historical financial analysis. Then they will submit to the Central Approval team for decision.

3.1.8.2. Final Credit/Loan Approval

Once the analyst and the relationship manager are satisfied with the contents of the loan approval document, they will present it to the credit officer for approval. Once again, there will likely be a few more rounds of edits to shape the deal into something that the credit officer is comfortable with. When the credit officer is comfortable with the terms, he/ she will sign it along with the relationship manager. Once the credit officer has approved the request, the relationship manager will send a term sheet to the borrower, outlining the bank approved structure of the proposed loan. The Borrower may attempt to negotiate certain points of the deal, but usually doesn't have a lot of room to work with. If the Borrower accepts the terms of the loan, they will sign the term sheet and the bank will issue a commitment letter.

Once the term sheet has been signed and the bank has issued a commitment, the loan will be routed to either the bank's internal loan operations area or to an attorney for the preparation of the loan documents. After the loan documents are prepared and the borrower (or the borrower's attorney) has reviewed them, the bank and the borrower will meet to sign them and the loan process is complete and credit release will take place

3.1.9. Conceptual Modeling

Once the required knowledge is acquired from credit applicant cases, loan experts and other relevant documents, the next step is modeling the knowledge. The knowledge modeling step involves organizing and structuring of the knowledge gathered during knowledge acquisition. This activity provides an implementation independent specification of the knowledge to be represented in the knowledge base. Knowledge modeling is the concept of representing information and the logic for the purpose of capturing, sharing and processing knowledge to simulate intelligence (Makhfi, 2011). Here, the basic concepts that tell the main activities and decisions are made to solve cases in the domain are modeled. Conceptual modeling is a crucial step in the knowledge acquisition process so as to understand well the problem domain and to prepare the knowledge representation phase.

Knowledge acquired through different knowledge acquisition techniques can be modeled with decision tree and hierarchical tree structure. Decision trees are produced by algorithms that identify various ways of splitting a data set into branch like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The hierarchical tree diagram provides the analyst with an effective visual condensation of the clustering results. The hierarchical tree diagram is one of commonly used methods of determining the number of clusters. It is also useful in spotting outliers, as these will appear as one member clusters that are joined later in the clustering process. The numbers at the top and bottom of the hierarchical tree diagram represent equally spaced values of the criterion function. It gives a pictorial representation of the criterion function information (Chen et al., 2003).

For this study, decision tree structure was used to represent knowledge modeling. Decision tree structure can easily model concepts and clearly explains the concepts in the problem area at hand. It models the knowledge in the tree structure manner. This model starts from the main concept at the highest level of the tree and other sub concepts that can affect or affected by the highest level concept put next to down ward in the tree (Makfi, 2011).

3.1.9.1. Credit approval Process and Decision Tree Structure Logical View

Credit Applications should summaries the results of the RMs risk assessment and include, as a minimum, the following details like Amount and type of loan(s) proposed, Purpose of loans, Loan Structure, Repayment Schedule, Interest, Security Arrangements etc. The decision tree depicted in figure 4 shows the different levels of decisions that credit approval team members use during Credit approval process

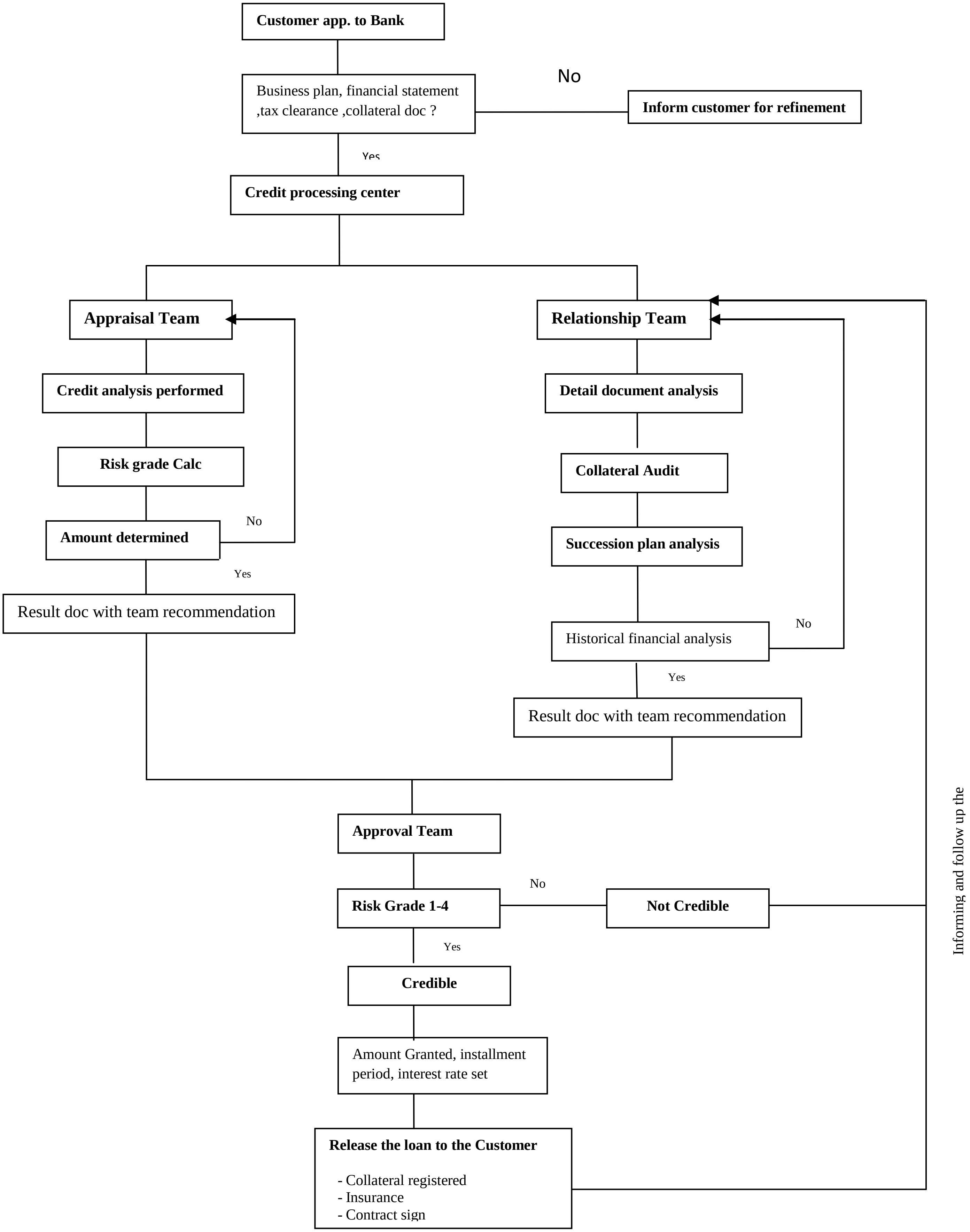

Customer app. to Bank
Business plan, financial statement ,tax clearance ,collateral doc ?
No
Inform customer for refinement
Yes
Credit processing center
Appraisal Team
Credit analysis performed
Risk grade Calc
Amount determined
No
Yes
Result doc with team recommendation
Relationship Team
Detail document analysis
Collateral Audit
Succession plan analysis
Historical financial analysis
No
Yes
Result doc with team recommendation
Approval Team
Risk Grade 1-4
No
Not Credible
Yes
Credible
Amount Granted, installment period, interest rate set
Release the loan to the Customer
- Collateral registered
- Insurance
- Contract sign
Informing and follow up the

Figure 3. 1 Credit approval process of CBE

As shown in the figure 3.1, the process begins when a customer applies for a loan by providing an application which comprises basic details of information like business plan, finical statement, and purpose of loan, loan amount and collateral document. Then the CRM begins preliminary discussions with the customer about the loan. These discussions typically include the amount, term, and rate of the loan. Then the CRM sends the application to the credit processing center. Once the application reached in central processing center, it directly sends to Appraisal and Relationship team members for separate and detail analysis. Then each team starts its own analysis and sends the final output which comprises results document and team member's recommendation to the main Approval Team. Here with in line with the two mentioned team members the credit officer will make decisions about the credibility of the customer or not. Once again, there will likely be a few more rounds of edits to shape the deal into something that the credit officer is comfortable with. When the credit officer is comfortable with the terms, he/ she will sign it along with the relationship manager.

Once the term sheet has been signed and the bank has issued a commitment, the loan will be routed to either the bank's internal loan operations area or to an attorney for the preparation of the loan documents. After the loan documents are prepared and the borrower (or the borrower's attorney) has reviewed them, the bank and the borrower will meet to sign them and the loan process is complete and credit release will take place. The final follow up and future communication to the customer will be left for CRM team members.

3.2. Knowledge Acquired from Data Mining

Knowledge acquisition is a process of identifying the knowledge, representing the knowledge in a proper format, structuring the knowledge, and transferring the knowledge to a machine. This process can be affected by the roles of the knowledge engineer, the expert and the end user (Bhambr, 2011). In addition, knowledge can also be acquired from large collection of dataset by using knowledge discovery tools. This type of knowledge is called hidden knowledge. Traditional knowledge acquisition techniques including on-site observation, protocol analysis, structured and unstructured interviewing and others can be used. As stated by Cornelius (2005), there are significant problems with each of these techniques. None of them guarantees consistency and integrity in the knowledge base. Some of the problems mentioned are: they are labor intensive, expensive to implement, expert conservatism and unwarranted biases.

Due to the aforementioned problems knowledge engineers look for other means to expand rule set and verify the rules already in the knowledge base. As a result Mihaela (2006) and Charles & Duminda (2002) stressed the need for developing automated techniques for knowledge acquisition. Considering the limitations mentioned above for acquiring knowledge from experts using traditional knowledge acquisition techniques, the researcher used data mining techniques for the development of the case based system. In this study, which focuses on designing and developing case base system for credit approval decision making, hence data mining specially classification algorithms are employed to generate cases for the case base.

Nowadays, data stored in banks databases are growing in an increasingly rapid way due to this tendency for data mining application in financial sectors today is great, because financial organizations today are capable of generating and collecting a large amounts of data. This increase in volume of data requires automatic way for these data to be extracted when needed. With the use of data mining techniques it is possible to extract interesting and useful knowledge and these knowledge can be used by experts for efficient and enhanced decision making process. In addition to this, Data mining tools can be very useful to control limitations of people such as subjectivity or error due to fatigue, and to provide indications for the decision-making processes specially on such error prone areas like Banks (Shapiro, 2001).

In addition, Knowledge acquisition is a complex and time-consuming stage during case based system development (Medi, 2008). For case generation and model building, classifier algorithms such as J48, PART, and naïve Bayes are employed and their result is compared to generate best rules and representative model for the case based system. Knowledge representation schemes such as frames, cases, semantic rules, and rule-based systems exist to represent knowledge (Charles & Duminda, 2002).

The data for this study have been collected from Commercial Bank of Ethiopia. Before 2006 the Bank uses manual way of keeping loan records within hard copy files for all districts and Branches. But after the Licensing and Supervision of banking business directive number SBB/1/94 issued by the National Bank of Ethiopia and come into effect as of the 15th day of April 2006 the bank started handling customer cases in a data base and reporting loan cases for National Bank of Ethiopia (NBE, 2000).

These directives are issued by the National Bank of Ethiopia pursuant to the authority vested in by article 41 of the Monetary and Banking Proclamation No. 83/1994 and article 36 of the Licensing and Supervision of Banking Business Proclamation No. 84/1994. The purpose of this Directive is to provide uniform guidelines to all banks to assure two important points. First, Loans or advances are regularly reviewed and classified in a manner consistent with regulatory standards; and secondly, Loans or advances which are not performing in accordance with contractual repayment terms are recognized and reported as past due in a manner consistent with regulatory standards.

Due to this, Banks are encouraged to obtain credit information from the Credit Information Center on prospective borrowers irrespective of the size of the loan. However, from the effective date of these directives, no bank shall extend new, or renew, reschedule or refinance existing, loans or advances equivalent to, or above, Birr 200,000 (two hundred thousand) without first obtaining credit information on borrowers from the Credit Information Center.

As results of this, all banks in Ethiopia start preparing and exchanging their customer loan files to National Bank of Ethiopia secure Data Base. The data set the researcher used are data from September 2007- Augest 2012 which are prepared with the aim of reporting the CBE credit customers data to NBE. The datasets include different attributes and status reports of the loan cases carried out by CBE in the upper mentioned period. The researcher used loan cases which have a known final status about their final credibility within the three (3) ranges: namely, Approved _good, Non-Performing and Denied.

Approved good:- Loans which have more than 24 months payment and no more than two (2) payments between 30 and 59 days.

Non performing:- Loans whose credit quality has deteriorated such that full collection of principal and/or interest in accordance with the contractual repayment terms of the loan or advance is in question. And loans which has payment history greater than 60 days due date.

Denied:- Application denied by the bank officers before approval.

The following table 3.1 summarizes the number of attributes and number of records that the researcher has collected from the bank database. The datasets collected from CBE within different years have the same number of attributes since they are prepared for the sake of reports to NBE central database with similar report formats.

Year	**No of Attributes**	**No of Records**
2007	**18**	246
2008	**18**	279
2009	**18**	340
2010	**18**	230
2011	**18**	250
2012	**18**	173
	Total	1518

Table 3. 1 Number of Attributes and Records

3.2.1. Data Preprocessing

Today’s real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their huge size (often several gigabytes or more) and origin from multiple, heterogeneous sources (Zhuang *et al., 2009)*. Therefore, prior to giving the data to a data mining tool, preprocessing of the data is necessary. Preprocessing the data includes multiple steps to assure the highest possible data quality, thus efforts are made to detect and remove errors, resolve data redundancies, and taking into account of the patient privacy, to remove patient identifiers (Inderpal, 2013). Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining (Zhuang, *et al., 2009)*.

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store. Data transformations, such as normalization, may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features. These techniques are not mutually exclusive; they may work together for a better data quality. Data mining requires access to data. The data may be represented as volumes of records in several database files or the data may contain only a few hundred records in a single file. According to (Inderpal, 2013), there are three common ways to access data for data mining:

i. Data can be accessed from a data warehouse,
ii. Data can be accessed from a database or
iii. Data can be accessed from a flat file or spreadsheet. In our case we used dataset from a flat file or a spreadsheet. In our case the researcher used dataset from a flat file or a spreadsheet.

Since the researcher uses a dataset which is preprocessed for the sake of reporting the customer credit reports to NBE central credit database, it passes most of the pre-processing steps. But the researcher performs other preprocessing activities to make the data more suitable for data mining.

Data formatting

Like any other software, WEKA needs data to be prepared in some formats and file types. The datasets provided to this software were prepared in a format that is acceptable for WEKA software. WEKA accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format (a file name with an extension of ARFF i.e. FileName.arff). At first the integrated dataset was in an excel file format. To feed the final dataset into the WEKA DM software the file is changed into other file format. The excel file was first changed into a comma delimited (CSV) file format. After changing the dataset into a CSV format the next step was opening the file with the WEKA DM software.

Data cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data (Zhuang *et al., 2009)*. Anticipating that data will be 100% complete and error free is unrealistic when working with financial data which collected in complex financial systems. Cleaning the data is proved a nontrivial and tedious task. Data error identification is both an automated and a manual process, and required an iterative procedure that drew upon expertise from the loan experts as well as statistical experts and the database administrators (Inderpal, 2013).

The researcher cleaned the data that has been collected from CBE head office separately according to the year. Datasets after the year 2012 are not included the final loan payment status are incomplete, as a result the researcher excluded the entire data from the data set. Since the data for all years have been recorded for each month separately, first the researcher integrated these data into one to make the data cleaning process more convenient and easy. Therefore, the data that have been recorded for each month within the years 2007-2012 sheets have integrated into one sheet and become ready for the next preprocessing step. In line with this some attributes which are believed confidential or irrelevant for the decision making process by the experts are removed prior to the data preprocessing task. Those attributes with their respective reasons are mentioned below in table 3.2. Beside this missing values of attributes are handled as described with table 3.3.

Removed attributes		
No	**Attribute name**	**Reason**
1	Name of Applicant	The data is confidential
2	Date Application Received and Date Application approved	The data is irrelevant for the decision making
3	Customers Business Address	The data is confidential.
4	Telephone	The data is confidential.
5	Nationality	The data is irrelevant for the decision making (all cases obtained has the same nationality- all are Ethiopian).

Table 3. 2 Removed attributes

Handling Missing Values				
Attribute name	**No of missing values**	**Data type**	**Method**	**Replaced with**
Length in Current address	13	Numeric	Mean	3
Age	3	Numeric	Mean	35

Table 3. 3 Missing values

Since the attribute values for "Account Balance" for 2007 and 2008 data and “Current Account Balance" attributes for the remaining years are similar, the researcher removed the " Account Balance " attribute from the record to avoid redundant data. In addition the bank uses “purpose” and “purpose of the credit” interchangeably in the dataset, the researcher uses “purpose” for connivance.

There are two attributes (Age (years) and Age) in the data which have the same data and due to that the researcher removed the former one to avoid redundancy. The organization has used different codes for " Current Account Balance ", " Payment Status of Previous Credit", " Collateral type ",” Length of current Address”, “Marital status”, and “Final Status”. The researcher used those codes since the experts are familiar with the codes in their daily activity.

Data Transformation

Data Transformation techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to a concise and easy to use knowledge level representation of mining results (Inderpal, 2013).

The data entries for the occupation attributes vary from customer to customer as a result the researcher merged them in to two categories namely private and governmental according to the description given within the data set for ease of use.

Since the attribute values of “Current Account Balance” is continuous and various, the researcher use data transformation and replace the actual data with the result to make the data more suitable for data mining according to the following ranges:

- No running account
- Between 1 and 100,000 Birr
- More than 100,000 Birr

Dataset Description

The collected data set have a total of 18 attributes. The table 3.4 below shows the data set description of CBE data.

No	Attributes	Description	Values	Type
1.	Current account balance	Current account balance of the customer in the Bank	No running account 1<=…<100,000 birr More than100,000 birr	Numeric
2.	Duration in current address	Duration in current address of the customer in years	Unknown, <=1year, 1-4 years, 5-10 years	Numeric
3.	Age	Age of the customer	Numeric values	Numeric
4.	Sex	Gender characteristics of the customer	f- female m-male	Nominal
5.	Marital status	Marital status of customer	Single Divorced Married	Numeric
6.	Concurrent credits	Further running credits	At this bank At other banks No further running credits	Nominal
7.	Occupation	Job type of the customer	Pi-private Gov't-government	Nominal
8.	No of credits at this Bank	Total number of credit at this Bank (this shows the	Numeric value	Numeric
9.	No of dependents	Number of persons under his full support and control	Numeric value	Numeric
10.	Payment status of previous credit	Customers credit re payment status	No previous credit current special mention Sub standard Dough full	Nominal
11.	Credit amount	The amount of the credit in birr	Numeric value	Numeric

12.	Purpose	Purpose of the credit (reason for the credit request)	Working capital, fixed asset, domestic trade and services, agriculture,	Nominal
13.	Duration of credit	Duration of credit in months	Numeric value	Numeric
14.	Installment period	Installment period of the credit	Monthly, Quarterly Semi-annually, Annually	Numeric
15.	Collateral type	The collateral the customer provides as a guarantor for the credit	Building/house, vehicle, Machinery, Bond, Farm, Other	Nominal
16.	Succession plan	The succession plan readiness of the customer	Yes or No	Nominal
17.	Creditability	Whether the customer is credible or not	Yes or No	Nominal
18.	Final status	Repayment status of the Customer	Approved good, Non-performing or Denied	Nominal

Table 3. 4 Data set description

3.2.2. Attribute Selection

In processing finacial data, choosing the optimal subset of features is such important, not only to reduce the processing cost but also to improve the usefulness of the model built from the selected data (Inderpal, 2013). The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit of reducing the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

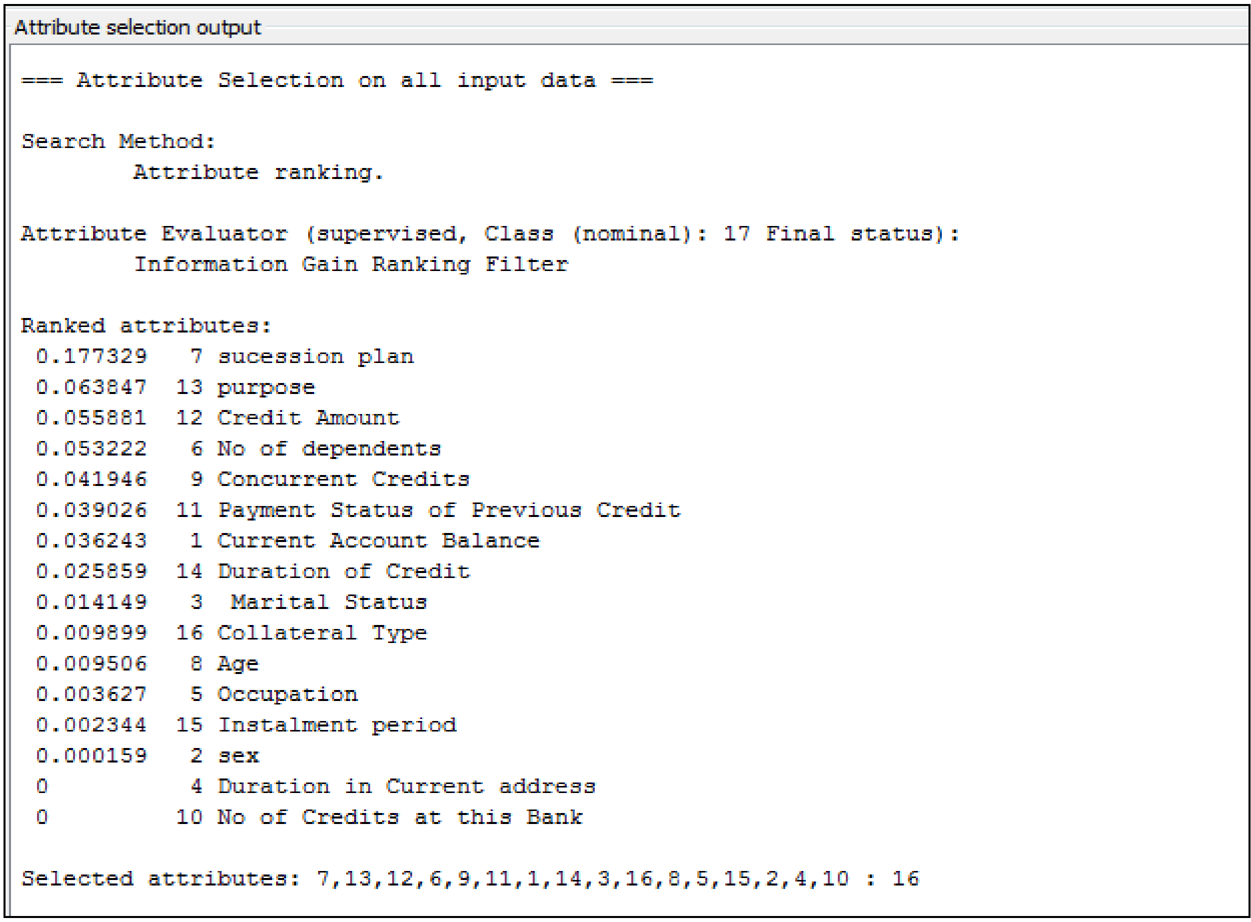
To select the best attributes for data mining, the researcher uses information gain method which exists in WEKA data mining tool. The attribute with the highest information gain is selected as the splitting attribute. This attribute minimizes the information needed to classify the instances in the resulting partitions and reflects the least impurity in these partitions. Entropy (impurity) is used to measure the information content of the attributes. High entropy means the attribute is from a uniform distribution where as low entropy means the attribute is from a varied distribution.

Before calculating the information gain of the attributes the researcher discussed with domain experts to select the most significant attributes for decision making. According to experts opinion all 18 attributes have their own value during credit approval in different circumstances. But they select 6 attributes which are believed to be more significant for decision making than the others. Beside this all experts agree those confidential attributes removed prior to getting the data were not significant for such decisions in the real environment. But customers profile is confidential and secrete according to the banks customer guidelines those attributes were removed. Attributes which are believed to be more significant by experts listed here below in the Table 3.5.

S.No	Attribute Name	Value	Description
1.	Creditability	Nominal	Describes of the Customer Credibility result during at the time of application.
2.	Credit Amount	Nominal	Describes the Credit Amount the Customer asks
3.	Purpose	Nominal	Describes purpose of the credit
4.	Collateral Types	Nominal	Describes collateral types the customer offer
5.	Concurrent credits	Nominal	Describes whether the customer has dependents or
6.	Payment status of previous credit	Nominal	Describes payment status of previous credit of the customer

Table 3. 5 Attributes selected by experts

Based on expert's prior feedback about the significant of the all attributes for the credit approval process, the researcher used all of them except the solution attribute to computer the information gain within the dataset. Accordingly the following result is obtained.



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 Final status):
        Information Gain Ranking Filter

Ranked attributes:
 0.177329    7 sucession plan
 0.063847   13 purpose
 0.055881   12 Credit Amount
 0.053222    6 No of dependents
 0.041946    9 Concurrent Credits
 0.039026   11 Payment Status of Previous Credit
 0.036243    1 Current Account Balance
 0.025859   14 Duration of Credit
 0.014149    3  Marital Status
 0.009899   16 Collateral Type
 0.009506    8 Age
 0.003627    5 Occupation
 0.002344   15 Instalment period
 0.000159    2 sex
 0           4 Duration in Current address
 0          10 No of Credits at this Bank

Selected attributes: 7,13,12,6,9,11,1,14,3,16,8,5,15,2,4,10 : 16
```

Figure 3. 2 Information Gain Result for Attribute Selection

As indicated from information gain result of attribute selection, attributes like succession plan, Purpose, Credit amount, Number of dependents, Concurrent credits, Payment status of previous credits, Current account balance, Duration of credit, Marital status, collateral type, Age, occupation and Installment period got a respective weight based on the relationship within the dataset. But information gain result for Duration in current address and No of credits at this bank results null.

The researcher again discuses' with experts after the information gain results were obtained. As a result the following suggestion comes from experts.

I. The obtained weight might help for future analysis during risk grade calculation since there are no such formal weighting mechanisms for each attributes still in the current environment.

II. Based on information gain result the 5 attributes which registers a better weight are Succession plan, Purpose, Credit amount, Number of dependants and concurrent credits. But the experts disagree with this except credit amount and concurrent credits and they select payment status of previous credit, collateral type, purpose, credit amount and concurrent credits as the 5 important attributes for decision making. But with respect to succession plan attribute, the experts agree it has its own impact on repayment status of the customer based on their experience. In line with this, experts also agree that number of dependants have direct relation on creditability in some situations and it also has indirect relationship with succession plan.

III. According to the information gain result attributes namely Duration in current address and Number of credits at this bank obtained null. According to experts suggestion both attributes have an implication with respect to Character analysis which is one of the 5C's for credit analysis. Due to this fact the researcher forced to use all attributes for the data mining task.

3.2.3. Data Mining

Data mining refers to the application of algorithms for extracting patterns from data. Data mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/ evaluation step of the KDD process (Inderpal, 2013). The objective of this step is to apply three classification technique algorithms on Bank data set which have been collected from CBE and develop a model that can predict the credibility status of the customer so that to use the model for case based development. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes (Bhambri, 2011).

Classification is form of data analysis that can be used to extract models describing important data classes or to predict future data trends and classification predicts categorical (discrete, unordered) labels (Asghar &Iqbal, 2009).

3.2.3.1. Experimental setting

A total of three experiments aiming at building predictive models are undertaken. The sampled data set contains 1518 instances. The data set contains 18 attributes and all of them are involved in all experiments. In addition after undertaking a number of experiments, default value of parameters is taken into consideration for each classifier algorithm since it allows achieving better accuracy compared to modifying the default parameters values.

The researcher conducted three experiments for three classification algorithms: namely, J48 pruned, PART and naiveBayes.

Before conducting the experiment, the researcher split the data sets for training and testing. As results 10 % of the data set (152 instances) was used for testing and the remaining 1366 instances for training. The data set has three classes: namely, Approved_good, Non- performing and Denied.

3.2.3.1. Experiment #1 using J48 Pruned

Decision tree is a graphical representation of the relations that exist between the data in the database. It is used for data classification. The result is displayed as a tree, hence the name of this technique. Decision trees are mainly used in the classification and prediction. It is a simple and a powerful way of representing knowledge.

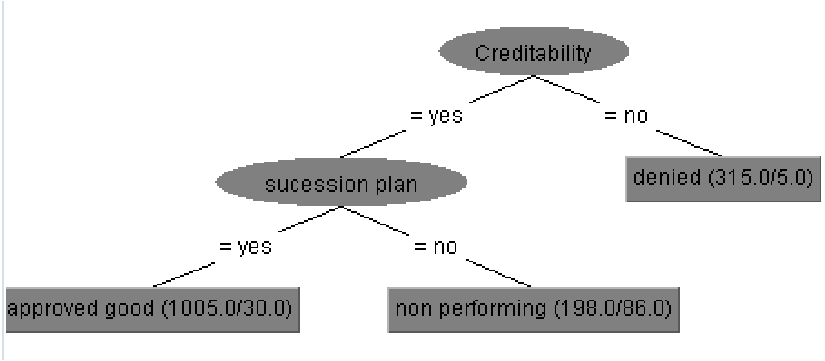


Figure 3. 3 Sample decision tree

The models obtained from the decision tree are represented as a tree structure. The instances are classified by sorting them down the tree from the root node to some leaf node. The nodes are branching based on if-then condition (Boris & Milan, 2012). This experiment conducted under percentage split test option with 90% of the data set for training and the remaining for testing with default parameters of WEKA and the algorithm generates a model as a decision tree with 38 Number of Leaves and 59 Size of the tree. As shown in table 3.6 Correctly Classified Instances are 1233 which means 90.2635 % and Incorrectly Classified Instances are 133 which means 9.7365 %from Total Number of Instances of 1366.

Confusion Matrix			
A	B	C	Classified as
952	0	4	a = Approved good
127	0	1	b = Non performing
1	0	281	c = Denied

Table 3. 6 Confusion Matrix for J48 decision tree

One of the compulsory steps of KDD methodology next to building classifier is evaluation of the model. Accordingly, the performance of the model has been evaluated based on the following criteria including performance accuracy, confusion matrix value, and True Positive rate and False Negative rate, Number of leaves and size of the tree generated and ROC curves and execution time. As shown in Table 3.7 the experimentation has performed in 10 percentage split test option.

Model Characteristics	**Experiment using J48**
Test option	90/10 % split
Pruned	Yes
Accuracy	90.2635 %
Time Taken	0.08seconds
Size of trees	59
AV.TPR (%)	0.903
AV.FPR (%)	0.219
AV.PR	0.82
AV.RR	0.903
AV.ROC	0.852
CCI	1233
ICI	133

Table 3. 7 Summery of J48 decision tree experimental result

Key: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate,

The result of **J48** based on correctly classified instance (out of 1366, 1233 instances are correctly classified), performance registered 90.2635 % accuracy and TP rate of 90.3% with 0.08 sec time taken for building model . The average ROC curve performance measure indicates J48 had 85.2 % performance

3.2.3.2. Experiment #2 using PART rule induction

PART is a rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form IF condition THEN conclusion. The "IF"-part (or left-hand side) of a rule is known as the rule antecedent or pre condition. The "THEN"-part (or right-hand side) is the rule consequent (Chen, 2009).

This experiment conducted under percentage splits technique using 90 % of instances for training and the remaining for testing with default parameters of WEKA and the algorithm generates a model with 83 rules. As shown in table 3.8 Correctly Classified Instances are 1193 which means 87.3353 % and Incorrectly Classified Instances are 173 which is 12.6647 % from Total Number of Instances of 1366.

Confusion Matrix			
A	B	C	Classified as
895	57	4	a = Approved good
110	17	1	b = Non performing
1	0	281	c = Denied

Table 3. 8 Confusion Matrix for PART rule induction

Accordingly, the performance of the model has been evaluated based on the same criterion that is applied for J48 classifier. As shown in Table 3.9 the experimentation has performed in 10 percentage split test option.

Model Characteristics	**Experiment using PART**
Test option	90/10 % split
Pruned	Yes
Accuracy	87.3353 %
Time Taken	1.55seconds
Number of Rules	208
AV.TPR (%)	0.873
AV.FPR (%)	0.195
AV.PR	0.847
AV.RR	0.873
AV.ROC	0.87
CCI	1193
ICI	173

Table 3. 9

Summery of PART experimental result

Key: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate,

The result of PART based on correctly classified instance (out of 1366, 1193 instances are correctly classified), performance registered 87.3353 % accuracy and TP rate of 87.3% with 1.55seconds time taken for building model. The average ROC curve performance measure indicates PART had 87 % performance.

3.2.3.3. Experiment #3 using naïve Bayes

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes (Ngai *et al.*, 2009; Han & Kamber 2006). This experiment conducted under percentage split test option with default parameters of WEKA. As shown in table 3.10 the algorithm generated a model in which Correctly Classified Instances are 1221 which means 89.3851 % and Incorrectly Classified Instances are 145 which means 10.6149 % from Total Number of Instances of 1366.

Confusion Matrix			
A	B	C	Classified as
909	40	7	a = Approved good
88	38	2	b = Non performing
3	5	274	c = Denied

Table 3. 10 Confusion Matrix for naïve Bayes classification algorithm

Accordingly, the performance of the model has been evaluated based on the same criterion that is applied for J48 and PART classifier. As shown in Table 3.11 the experimentation has performed in 10 percentage split test option.

Model Characteristics	**Experiment using naive Bayes**
Test option	90/10 % split
Pruned	Yes
Accuracy	89.3851 %
Time Taken	0.05 sec
Number of Rules	-
AV.TPR (%)	0.894
AV.FPR (%)	0.16
AV.PR	0.879
AV.RR	0.894
AV.ROC	0.932
CCI	1221
ICI	145

Table 3. 11 Summery of naïve Bayes experiment

Key: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate,

The result of naïve Bayes shows that based on correctly classified instance (out of 1366, 1221 instances are correctly classified), performance registered 89.3851 % accuracy and TP rate of 89.4% with 0.05 seconds time taken for building model . The average ROC curve performance measure indicates PART had 87 % performance.

3.2.4. Comparison of Classification algorithms

The researcher used objective interestingness evaluation methods. Objective interestingness measurement is generally based upon the inherent structure of mined patterns, i.e., the patterns' statistics like support or confidence (Pohle, 2003).

	Classification algorithms		
Objective evaluations	**J48 pruned**	**PART**	**naïve Bayes**
Correctly Classified Instances	90.2635 %	87.3353 %	89.3851 %
Incorrectly Classified Instances	9.7365%	12.6647 %	10.6149 %
Time taken per second	0.08 sec	1.55 sec	0.05 sec
TP Rate	0.903	0.903	0.894
FP Rate	0.219	0.219	0.16
Precision	0.82	0.82	0.879
Recall	0.903	0.903	0.894
F-Measure	0.859	0.859	0.884

Table 3. 12 Objective evaluation results

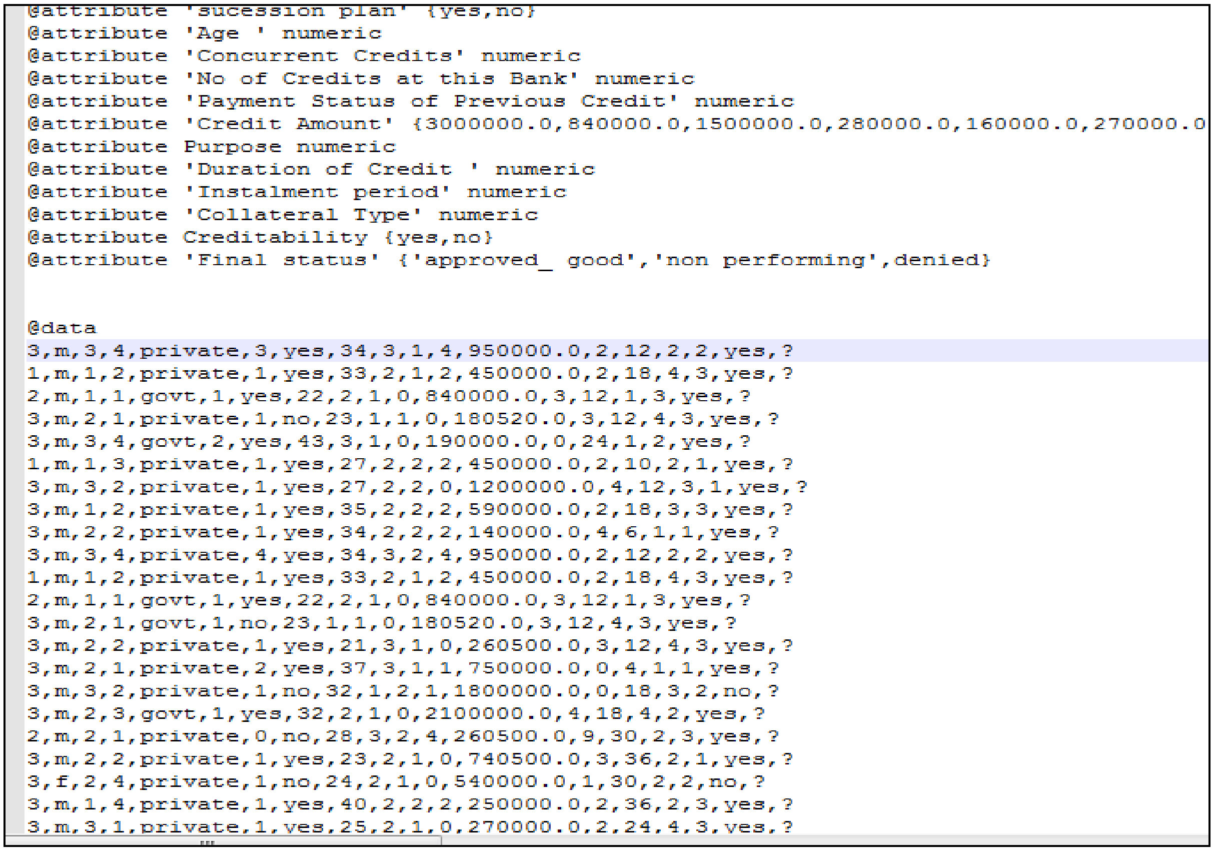
As we can observe from the table 3.12 above, J48 classification algorithm performed better in all objective interestingness evaluation methods than the PART and also performed better than naïve Bayes except a slight difference in running time. Based on the results obtained by objective interestingness evaluation methods result, the researcher decided to use J48 classification algorithm model for further use in the development of case base of the case base system because it registered better performance than PART and naïve Bayes Classification algorithms.

For this study a number of rules are generated by the J48 algorithm to identify an instance of the KDD dataset as Approved good, Non-performing and Denied. For that, most rules used combination of attributes and few of them used a single attribute with the respective values for attributes. Therefore generated rules evaluated in consultation with domain experts in the area of credit approval. As a result the following 11 rules are believed to be surprising by domain experts. The overall J48 classifier outputs are presented in the Appendix section.

S. no.	Rules
1	Succession plan = yes AND No of dependents <= 2 AND purpose = fixed asset AND Marital Status = married : approved good
2	Succession plan = yes AND purpose = fixed asset AND Marital Status AND Marital Status= single /Divorced : denied
3	succession plan = yes AND Current Account Balance <= 100,000 AND purpose = Manufacturing: denied
4	purpose = Domestic trade AND Status of Previous Credit = special mention : denied
5	Current Account Balance > 100,000 AND purpose = Domestic trade AND Payment Status of Previous Credit = substandard: approved good
6	succession plan = yes AND purpose = Agri AND Concurrent Credits = At this bank: approved good
7	purpose = working cap / domestic trade AND Duration in Current address <= 3: denied
8	succession plan = no AND purpose = fixed asset: non-performing
9	succession plan = no AND purpose = Manufacturing: denied
10	purpose = Domestic trade AND Duration of Credit > 15: non-performing
11	succession plan = no AND purpose = Agri / working cap : non-performing

Table 3. 13 Rules extracted by J48 Classification algorithm

After selecting the best model and classifier, the next task is to know how the model managed to classify the new instances. To test the model the researcher used 156 test instances which split from the main dataset before the data mining task takes place using percentage split technique (i.e 10 percent of the data set is used for testing). After setting the test data in the supplied test set, the results can be obtained from results list menu by using Visualize classifier errors menu indirectly by saving the results with .arff extension. As a result this file contains a copy of the new instances along with an additional column for the predicted value of "predicted Final status". Figure 3. 4 below depict the test set used for testing the prediction capability of the model created by J48.

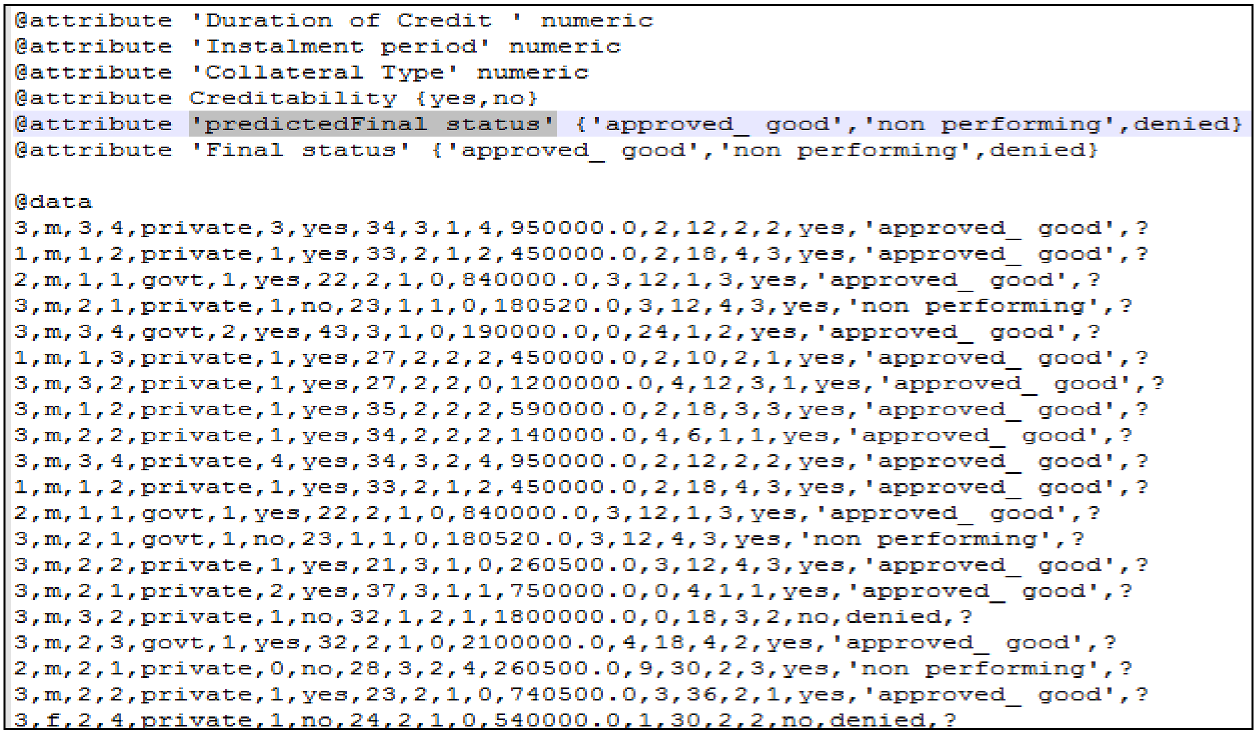


```
@attribute 'sucession plan' {yes,no}
@attribute 'Age ' numeric
@attribute 'Concurrent Credits' numeric
@attribute 'No of Credits at this Bank' numeric
@attribute 'Payment Status of Previous Credit' numeric
@attribute 'Credit Amount' {3000000.0,840000.0,1500000.0,280000.0,160000.0,270000.0
@attribute Purpose numeric
@attribute 'Duration of Credit ' numeric
@attribute 'Instalment period' numeric
@attribute 'Collateral Type' numeric
@attribute Creditability {yes,no}
@attribute 'Final status' {'approved_ good','non performing',denied}


@data
3,m,3,4,private,3,yes,34,3,1,4,950000.0,2,12,2,2,yes,?
1,m,1,2,private,1,yes,33,2,1,2,450000.0,2,18,4,3,yes,?
2,m,1,1,govt,1,yes,22,2,1,0,840000.0,3,12,1,3,yes,?
3,m,2,1,private,1,no,23,1,1,0,180520.0,3,12,4,3,yes,?
3,m,3,4,govt,2,yes,43,3,1,0,190000.0,0,24,1,2,yes,?
1,m,1,3,private,1,yes,27,2,2,2,450000.0,2,10,2,1,yes,?
3,m,3,2,private,1,yes,27,2,2,0,1200000.0,4,12,3,1,yes,?
3,m,1,2,private,1,yes,35,2,2,2,590000.0,2,18,3,3,yes,?
3,m,2,2,private,1,yes,34,2,2,2,140000.0,4,6,1,1,yes,?
3,m,3,4,private,4,yes,34,3,2,4,950000.0,2,12,2,2,yes,?
1,m,1,2,private,1,yes,33,2,1,2,450000.0,2,18,4,3,yes,?
2,m,1,1,govt,1,yes,22,2,1,0,840000.0,3,12,1,3,yes,?
3,m,2,1,govt,1,no,23,1,1,0,180520.0,3,12,4,3,yes,?
3,m,2,2,private,1,yes,21,3,1,0,260500.0,3,12,4,3,yes,?
3,m,2,1,private,2,yes,37,3,1,1,750000.0,0,4,1,1,yes,?
3,m,3,2,private,1,no,32,1,2,1,1800000.0,0,18,3,2,no,?
3,m,2,3,govt,1,yes,32,2,1,0,2100000.0,4,18,4,2,yes,?
2,m,2,1,private,0,no,28,3,2,4,260500.0,9,30,2,3,yes,?
3,m,2,2,private,1,yes,23,2,1,0,740500.0,3,36,2,1,yes,?
3,f,2,4,private,1,no,24,2,1,0,540000.0,1,30,2,2,no,?
3,m,1,4,private,1,yes,40,2,2,2,250000.0,2,36,2,3,yes,?
3,m,3,1,private,1,yes,25,2,1,0,270000.0,2,24,4,3,yes,?
```

Figure 3. 4 Test instances before prediction using J48 Classifier

Prediction result of J48 decision tree classification algorithm using the above test set is shown below in figure 3.5.



```
@attribute 'Duration of Credit ' numeric
@attribute 'Instalment period' numeric
@attribute 'Collateral Type' numeric
@attribute Creditability {yes,no}
@attribute 'predictedFinal status' {'approved_ good','non performing',denied}
@attribute 'Final status' {'approved_ good','non performing',denied}

@data
3,m,3,4,private,3,yes,34,3,1,4,950000.0,2,12,2,2,yes,'approved_ good',?
1,m,1,2,private,1,yes,33,2,1,2,450000.0,2,18,4,3,yes,'approved_ good',?
2,m,1,1,govt,1,yes,22,2,1,0,840000.0,3,12,1,3,yes,'approved_ good',?
3,m,2,1,private,1,no,23,1,1,0,180520.0,3,12,4,3,yes,'non performing',?
3,m,3,4,govt,2,yes,43,3,1,0,190000.0,0,24,1,2,yes,'approved_ good',?
1,m,1,3,private,1,yes,27,2,2,2,450000.0,2,10,2,1,yes,'approved_ good',?
3,m,3,2,private,1,yes,27,2,2,0,1200000.0,4,12,3,1,yes,'approved_ good',?
3,m,1,2,private,1,yes,35,2,2,2,590000.0,2,18,3,3,yes,'approved_ good',?
3,m,2,2,private,1,yes,34,2,2,2,140000.0,4,6,1,1,yes,'approved_ good',?
3,m,3,4,private,4,yes,34,3,2,4,950000.0,2,12,2,2,yes,'approved_ good',?
1,m,1,2,private,1,yes,33,2,1,2,450000.0,2,18,4,3,yes,'approved_ good',?
2,m,1,1,govt,1,yes,22,2,1,0,840000.0,3,12,1,3,yes,'approved_ good',?
3,m,2,1,govt,1,no,23,1,1,0,180520.0,3,12,4,3,yes,'non performing',?
3,m,2,2,private,1,yes,21,3,1,0,260500.0,3,12,4,3,yes,'approved_ good',?
3,m,2,1,private,2,yes,37,3,1,1,750000.0,0,4,1,1,yes,'approved_ good',?
3,m,3,2,private,1,no,32,1,2,1,1800000.0,0,18,3,2,no,denied,?
3,m,2,3,govt,1,yes,32,2,1,0,2100000.0,4,18,4,2,yes,'approved_ good',?
2,m,2,1,private,0,no,28,3,2,4,260500.0,9,30,2,3,yes,'non performing',?
3,m,2,2,private,1,yes,23,2,1,0,740500.0,3,36,2,1,yes,'approved_ good',?
3,f,2,4,private,1,no,24,2,1,0,540000.0,1,30,2,2,no,denied,?
```

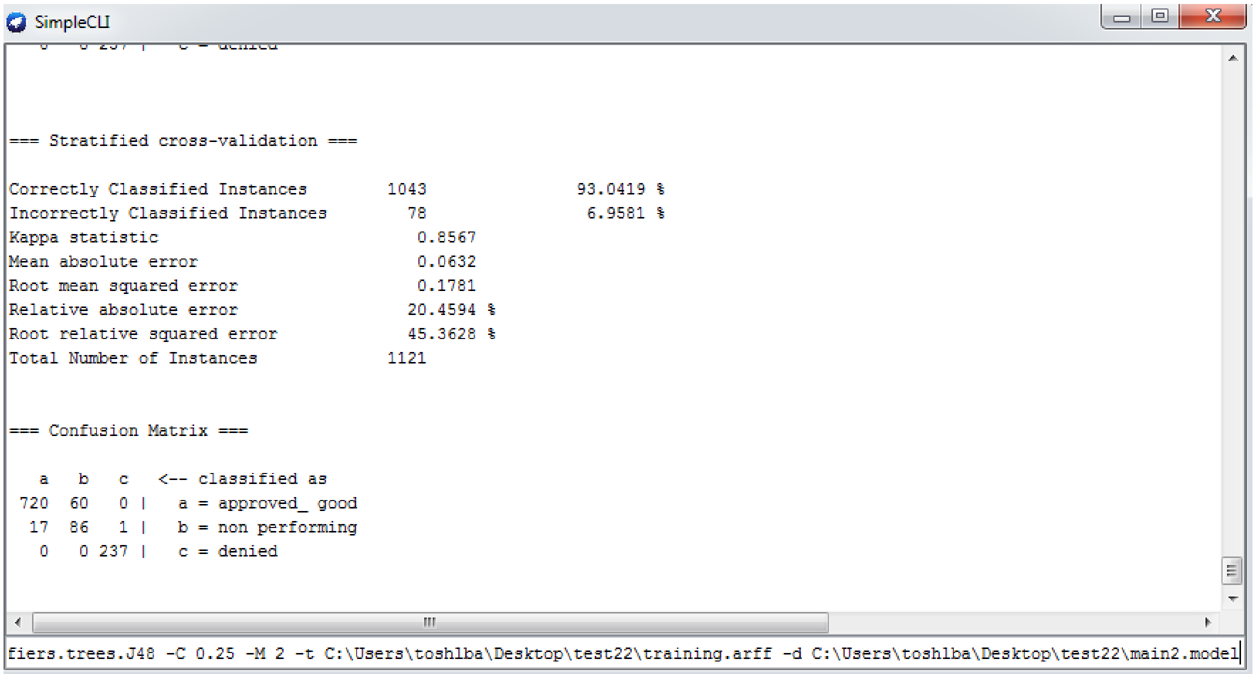
Figure 3. 5 Test instances after prediction using J48 Classifier

As shown in the figure above, the " predictedFinal status " value for each new instance is the last value before "?" which the actual " Final status " class value. For example, the predicted value of the "Final status" attribute for instance 1 is "Approved_good" according to the model, while the predicted class value for instance 4 is "non performing".

While the GUI version of WEKA is nice for visualizing the results and setting the parameters using forms, when it comes to building a classification (or predictions) model and then applying it to new instances, the most direct and flexible approach is to use the command line. In the main WEKA interface, click "Simple CLI" button to start the command line interface. The main command for generating the classification model is:

java WEKA.classifiers.trees.J48 -C 0.25 -M 2 -t C:\Users\toshlba\Desktop\test22\training.arff -d C:\Users\toshlba\Desktop\test22\main.model

The options -C 0.25 and -M 2 in the above command are the same options that is selected for J48 classifier in during GUI model building. The -t option in the command specifies that the next string is the full directory path to the training file (in this case "training.arff"). Finally, the -d option specifies the name (and location) where the model will be stored. After executing this command inside the "Simple CLI" interface, you should see the tree and stats about the model in the top window.



```
   0   0 237 |   c = denied



=== Stratified cross-validation ===

Correctly Classified Instances        1043               93.0419 %
Incorrectly Classified Instances        78                6.9581 %
Kappa statistic                          0.8567
Mean absolute error                      0.0632
Root mean squared error                  0.1781
Relative absolute error                 20.4594 %
Root relative squared error             45.3628 %
Total Number of Instances             1121


=== Confusion Matrix ===

   a   b   c   <-- classified as
 720  60   0 |   a = approved_ good
  17  86   1 |   b = non performing
   0   0 237 |   c = denied
```

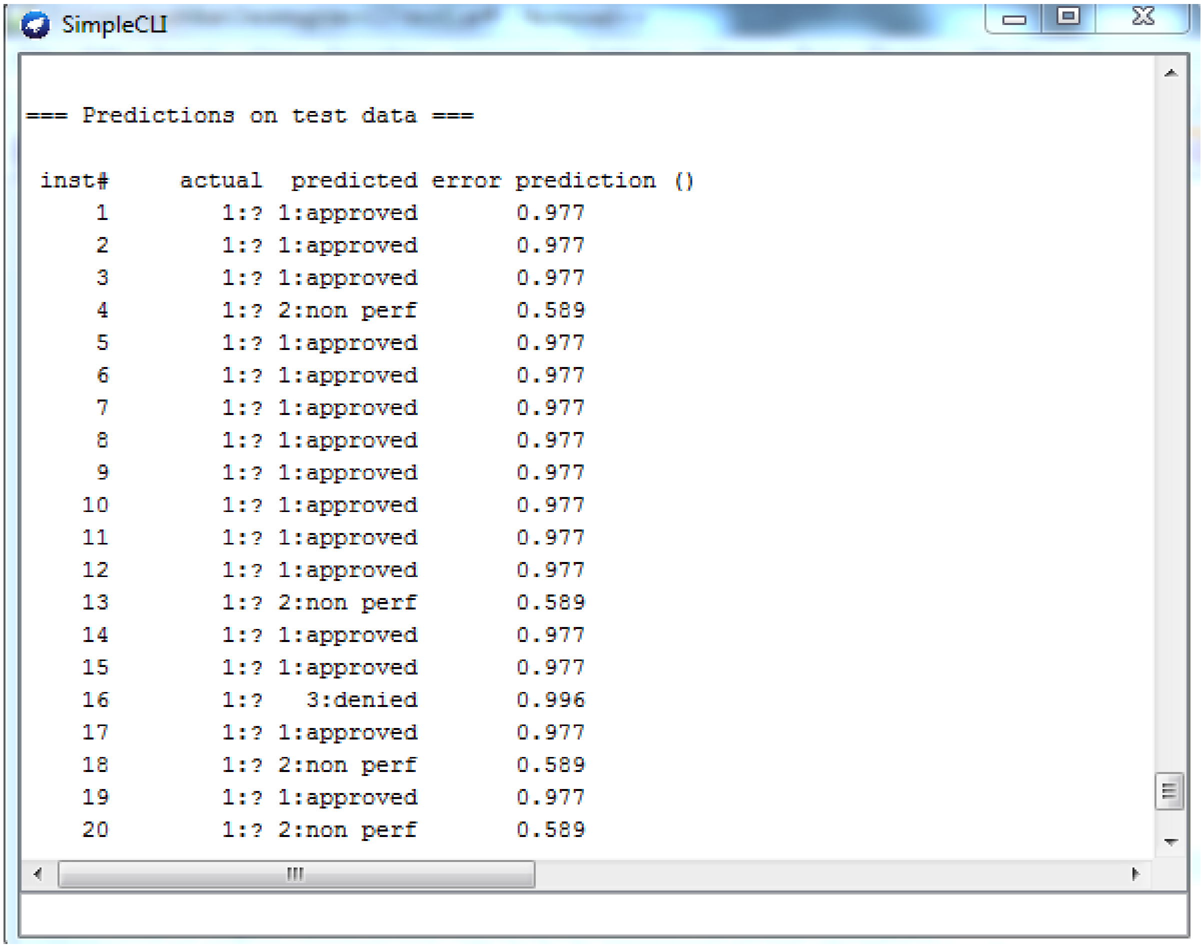
```
fiers.trees.J48 -C 0.25 -M 2 -t C:\Users\toshlba\Desktop\test22\training.arff -d C:\Users\toshlba\Desktop\test22\main2.model
```

Figure 3. 6 J48 model building using CLI feature of WEKA

Based on the above command, our classification model has been stored in the file "main.model" and placed in the directory specified. Now it is possible to apply this model to the new instances. The advantage of building a model and storing it is that it can be applied at any time to different sets of unclassified instances. The command for doing so is:

Java WEKA.classifiers.trees.J48 -p 18 -l C:\Users\toshlba\Desktop\test22\main2.model -T C:\Users\toshlba\Desktop\test22\test1.arff

In the above command, the option -p 18 indicates that the researcher want to predict a value for attribute number 18 (which is "final status"). The -l options specifies the directory path and name of the model file (this is what was created in the previous step). Finally, the -T option specifies the name (and path) of the test data. In our example, the test data is our new instances file "test1.arff").



```
=== Predictions on test data ===

 inst#     actual  predicted error prediction ()
     1        1:? 1:approved       0.977
     2        1:? 1:approved       0.977
     3        1:? 1:approved       0.977
     4        1:? 2:non perf       0.589
     5        1:? 1:approved       0.977
     6        1:? 1:approved       0.977
     7        1:? 1:approved       0.977
     8        1:? 1:approved       0.977
     9        1:? 1:approved       0.977
    10        1:? 1:approved       0.977
    11        1:? 1:approved       0.977
    12        1:? 1:approved       0.977
    13        1:? 2:non perf       0.589
    14        1:? 1:approved       0.977
    15        1:? 1:approved       0.977
    16        1:?   3:denied       0.996
    17        1:? 1:approved       0.977
    18        1:? 2:non perf       0.589
    19        1:? 1:approved       0.977
    20        1:? 2:non perf       0.589
```

Figure 3. 7 Sample CLI prediction results on test data using J48 classifier

The above output is preferable over the output derived from the GUI version on WEKA. First, this is a more direct approach which allows us to save the classification model. This model can be applied to new instance later without having to regenerate the model. Secondly (and more importantly), in contrast to the final output of the GUI version, in this case we have independent confidence (accuracy) values for each of the new instances. This means that we can focus only on those predictions with which are more confident. Since the data size the researcher obtained is small in size testing the accuracy of instances can help to get more reliable instances to use them as a knowledge base.

Due to this fact, the researcher used only those test instances whose predicted value has an accuracy of greater than 95% as an input for the prototype development.

CHAPTER FOUR

DESIGN AND IMPLEMENTATION OF THE PROTOTYPE

The design and implementation part of this section involves the actual development of a scaled down workable CBR system for credit approval decision making. Therefore, having all the necessary cases and the knowledge from automatic knowledge acquisition using Data mining and domain know how from experts, the next task is coding the knowledge into computer using appropriate and efficient knowledge representation methods. For this research, j COLIBR 1.1 CBR frame work is used to develop the prototype.

The retrieval algorithm used in this research is nearest neighbor retrieval algorithm. Because JCOLIBRI uses this algorithm for retrieval task. Nearest neighbor retrieval algorithm is also suitable when there are attributes which have numeric (continuous) value (Fang & Songdong, 2007).

4.1. Designing the Architecture of CBRCADM

The architecture of the CBRCADM system shown in figure 4.1 depicts how the prototype works during Credit approval decision making. As the new query (problem) is entered, the prototype of the system matches the new case to the solved case in the case base of the system by using similarity measurement. If relevant cases are found within the case base, then the prototype rank the relevant retrieved cases based on their global similarity. Next, the prototype proposes a solution. The proposed solution can be derived directly from a retrieved case that matches exactly or partially to the problem of the new case. But, using the proposed solutions directly may have a risk. Therefore, the user of the system should make an adaptation by altering the differences between the proposed case and the new case. In addition to adaptation, case inconsistencies are revised if the retrieved case is not the same as the new case. Finally, the revised solution is retained in the case base for future problem solving.

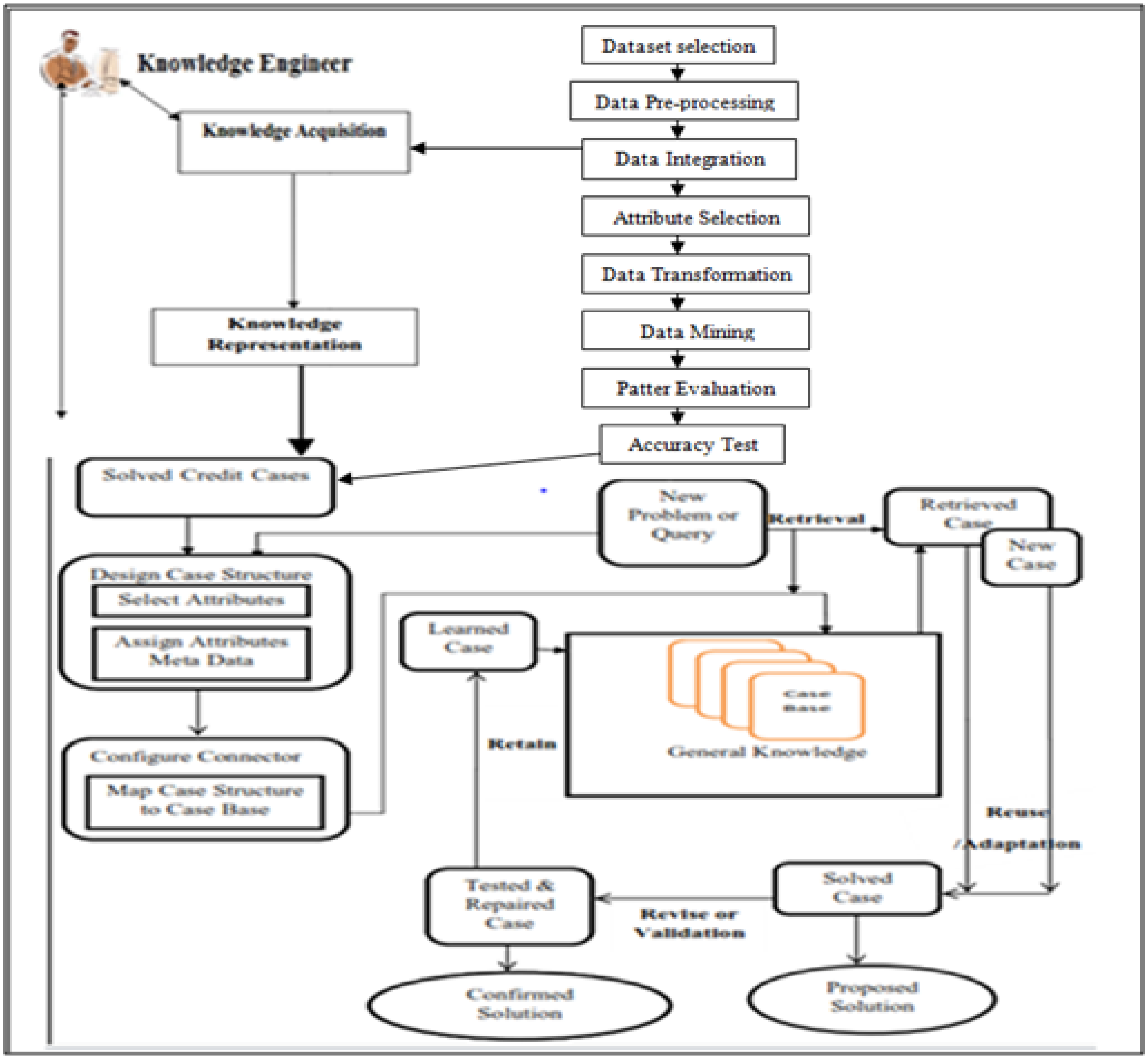


Figure 4. 1 CBRCADM Architecture (Adopted from Grimnes & Aadmot, 1996)

4.2. Case-based Reasoning System for Credit Approval Decision Making

The development of a reasonable simple CBR application already involves a number of steps, such as collecting cases and background knowledge, modeling a suitable case representation, defining an accurate similarity measure, implementing retrieval functionality, and implementing user interfaces (Stahl & Roth-Berghofe, 2008). In this study, the researcher used the main feature of JCOLIBRI to deliver the actual prototype. As Recio- Garcia, Diaz-Agudo & Gonzalez-Calero

(2008) presented JCOLIBRI has been constructed as a core module to offer the basic functionality for developing CBR application. Implementing a CBR application from scratch remains a time consuming software engineering process and requires a lot of specific experience beyond pure programming skills (Stahl & Roth-Berghofe, 2008). Therefore, using JCOLIBRI CBR framework minimizes the effort to develop an application by using other programming languages.

To run JCOLIBRI for the first time, click on the JCOLIBRI.bat file and it becomes ready for usage as shown in the following figure 4.2.

Figure 4. 2 Main Window of JCOLIBRI

Developing a CBR system is a complex task where many decisions must be made. In this study, the development of the CBR system for Credit Approval decision makings is divided into the following subsections which enable to achieve the objectives of the research.

4.2.1Building the Case Base

During setting up the objectives of this study, one of the objectives is collecting credit cases in order to build a case base and represent the cases using the appropriate case representation method. So, the researcher collected credit cases from Commercial Bank of Ethiopia head office and used Automatic knowledge Acquisition technique using data mining tool WEKA. As discussed in Chapter 3, the test instances that scores more than 95 % of accuracy are used to build Credit approval decision making CBR system that is used to assist by offering decision support to Credit analysts, Credit officers, CRM Mangers and other board members. All the acquired cases are stored as plaintext files in a feature-value representation format, as a result the .ARFF results of the test cases in to .txt format. The case base is presented as a plaintext comprising of *n* columns representing case attributes (A1, A2, A3, ..., An) and each *m* rows representing individual cases C ({C1, C2, C3, . ,.,Cm}) each attribute has a sequence of possible *k* values associated to each column attribute A={V1, V2, V3, ., Vk}. The reason for representing cases using feature-value representation is that this approach supports nearest neighbor retrieval algorithm and it represents cases in an easy way (Salem *et al.*, 2005; Bergmann *et al.*, 2005).

4.2.2. Case Representation

The case representation is made in a way that easily fit to JCOLIBRI. Designing of such a case structure helps easily define the features available in the case and to measure the similarity between existing and new cases. Hence, the overall application of this research is to retrieve similar cases from the case base that can guide future reasoning, problem solving and also transforming a solution retrieved in a solution appropriate to the current problems. The collections of cases are represented in the feature-value representation to make efficient retrieval process. This is done through case indexing process. Indexing refers to assigning indices to cases for retrieval and comparison of a query to the case base (Luzelschwah, 2007).

4.2.3Managing/Defining the Case Structure in JCOLIBRI

The acquired cases are saved in plaintext file format. Case attributes are succession plan, Credit Amount, Purpose, No of dependents, Concurrent Credits, Current Account Balance, Payment Status of Previous Credit, Duration of Credit, Collateral Type, Marital Status, Sex, age, Duration in Current address, Occupation, Succession plan, Creditability and Final status are the important attributes chosen from all attributes. These attributes have significant impact on Credit Approval process.

Most of the case attributes have Numeric values and a few attributes have Nominal values. Local similarity functions are used to compare simple attribute values. In this research, the following local similarity functions are used.

Equal: If you select equal local similarity for each attribute. Then your input and value of case base must be match. If value matches exactly then it will get result otherwise match failure.

Interval: When you select similarity interval and adjust interval value. Then, JCOLIBRI match value keeping in mind that interval. Exact value match is not compulsory in that type.

Global Similarity is linked with compound attributes and used to get similarity of collected attributes in unique similarity value. Global similarity used in this research is average similarity.

Average: It is a type of global similarity that considers the average of all attribute local similarity values. The algorithm works as follows (Watson & Marir, 1994; Salem et al., 2005; Henok, 2011).

Step 1: Find the local similarity of step for all attributes of the case which make up the case base

Step 2: Multiply the result of the local similarity of attributes with their corresponding attribute weight (importance value)

Step 3: Add the value of all attribute results of step 2

Step 4: Add all weights of attributes that represent the importance value of the attributes and multiply by the number of attributes

Step 5: Divide the result of step 3 by the result of step 4 and the result of this step is the global similarity that represents the degree of match of the old case with the new input case

4.3.1.1 Description of CBRCADM Case Attributes

Defining case structure in JCOLIBRI are done by using simple manage case structure window. It is very easy to define case structure with JCOLIBRI. Because it is simple to add attributes in description of case structure and set properties of attributes or metadata of attributes. Metadata of attributes are weight of attribute, data type of attribute and similarity function. During configuration of case structures, JCOLIBRI creates codes automatically and saved in xml file format.

Most significant attributes are set by declaring higher weight as compared to other weights. Based on attribute selection task using information gain attribute evaluator results Succession plan, purpose, Credit amount, Number of dependency concurrent credits, payment status of previous credits and collateral type have more weight than other attributes. In contrast attributes namely, duration in current address and number of credits at the bank obtained a very low information gain results as per the Data Mining results which discussed in Chapter 3.

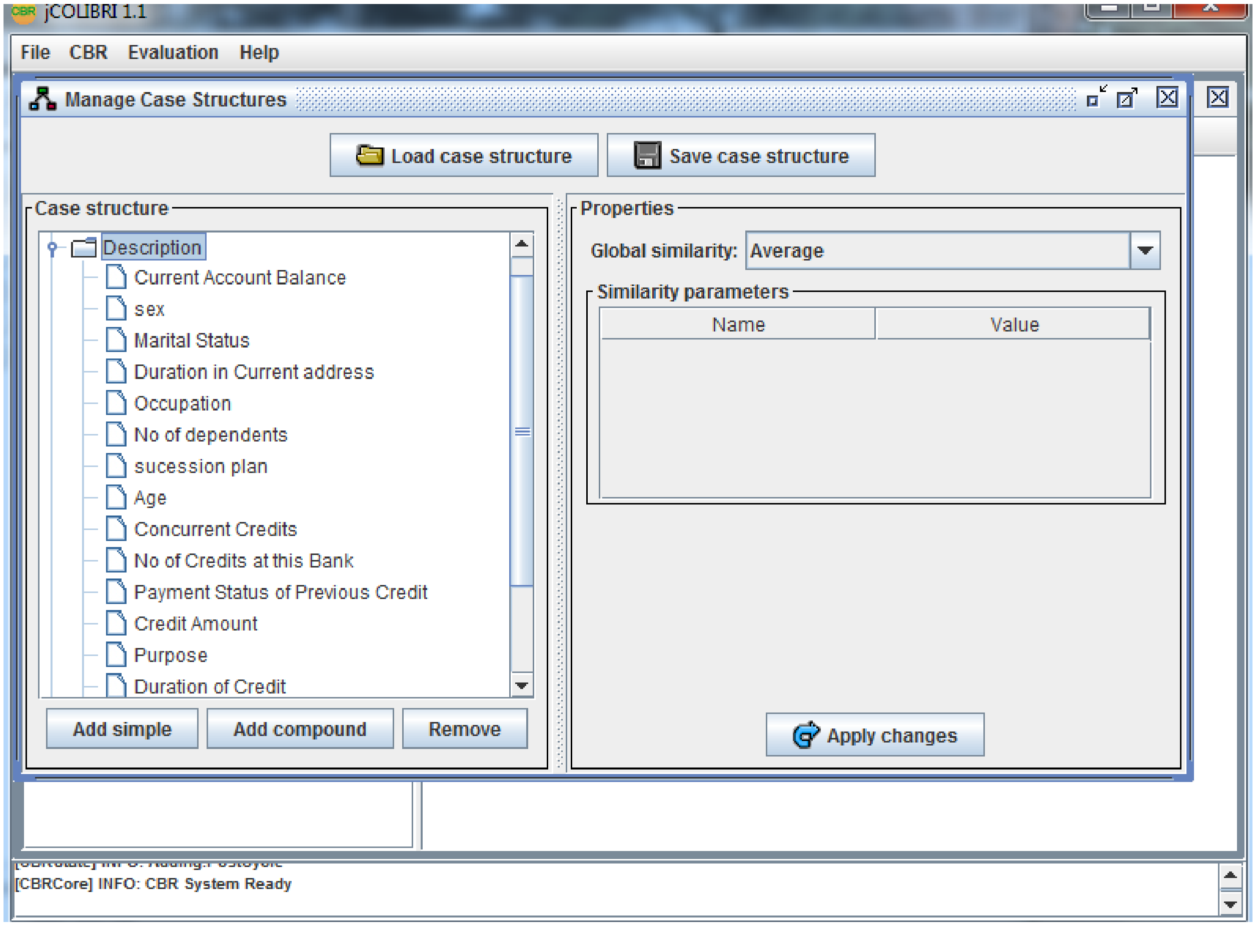
As a result for building CBRCADM the weights value for the attributes comes from attribute selection using information gain attribute evaluator and loan expert’s feedback on the results. As a result Purpose, Payment status of previous credit, Collateral Type, Concurrent Credits Succession plan and the two solution attributes got a weight of 1.0. The remaining attributes weights are given by discussing with experts on the credit domain and depict in table 4.1 below.

CBRCADM prototype case base has **16** description attributes and 2 solution attributes. Solution attribute is used after finding best selected cases and show the type of the decision making during loan approval.

The following table 4.1 shows the description of case attributes regarding name, data type, weights, local and global similarity.

Significant Attributes			
Attribute Name	**Data Type**	**Weight**	**Local Similarity**
Succession plan	Boolean	1.0	Equal
Age	Integer	0.1	Equal
Credit Amount	Integer	0.8	Equal
Purpose	Integer	1.0	Equal
Sex	String	0.4	Equal
Payment status of previous credit	Integer	1.0	Equal
No of dependents	Integer	1.0	Threshold
Duration in Current address	Integer	0.0	Threshold
Occupation	String	0.3	Equal
Concurrent Credits	Integer	1.0	Threshold
Credit Amount	Integer	0.6	Threshold
Duration of Credit	Integer	0.6	Equal
Installment period	Integer	0.0	Threshold
Collateral Type	Integer	0.8	Threshold
Current Account Balance	String	0.7	Equal
Marital Status	Integer	0.3	Equal
No of Credits at this Bank	Integer	0.3	Threshold
Solution attributes			**Global Similarity**
Creditability	String	1.0	EqualStringingnrecase
Final status	String	1.0	EqualStringingnrecase

Table 4. 1 Case Description



Description
- Current Account Balance
- sex
- Marital Status
- Duration in Current address
- Occupation
- No of dependents
- sucession plan
- Age
- Concurrent Credits
- No of Credits at this Bank
- Payment Status of Previous Credit
- Credit Amount
- Purpose
- Duration of Credit

Global similarity: Average

```
[CBRCore] INFO: CBR System Ready
```

Figure 4. 3 Configurations of Case Structure and Similarities

4.2.4 Managing Connectors

Once case structures are configured in JCOLIBRI, CBR systems must access the stored cases in an efficient way. JCOLIBRI splits the problem of case base management in two separate although related concerns: persistency mechanisms through connectors and in-memory organization.

Cases are often derived from legacy databases, thereby converting existing organizational resources into exploitable knowledge. To take advantage of these previously existing resources, facilitate intelligent access to existing information, and incorporate it as seed knowledge in the CBR system (the case base), JCOLIBRI offers a set of connectors to manage persistence of cases.

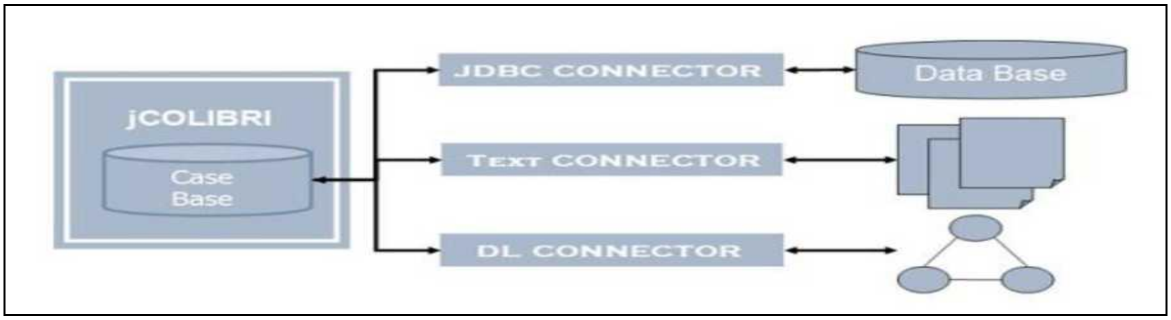
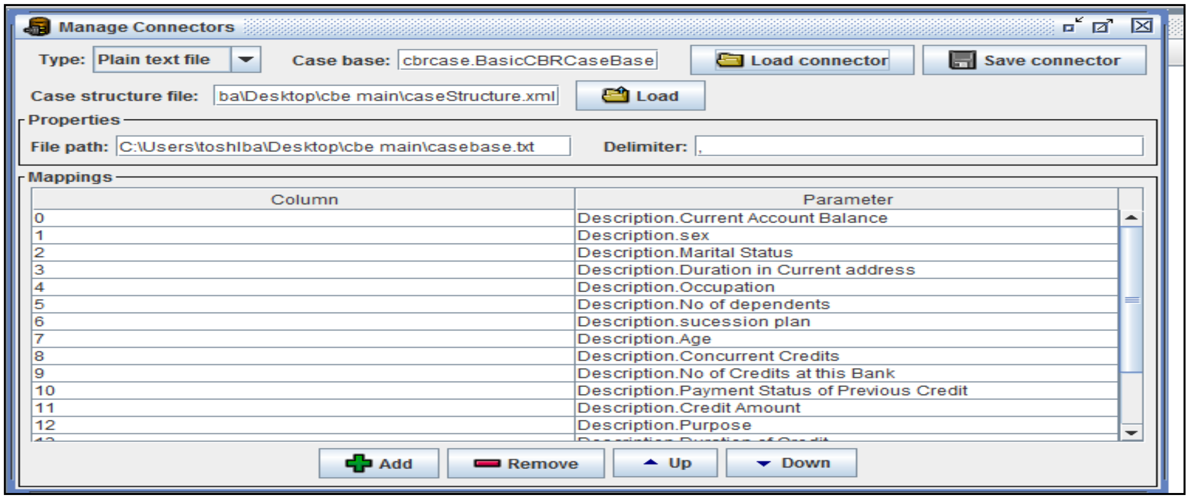


Figure 4. 4 JCOLIBRI Connector Schema

Connectors are objects that know how to access and retrieve cases from the storage media and return those cases to the CBR system in a uniform way. Therefore connectors provide an abstraction mechanism that allows users to load cases from different storage sources in a transparent way. As shown in figure 4.4, JCOLIBRI includes connectors that work with plain text files, relational databases and Description Logics systems. For the implementation CBRCADM prototype, the researcher used plaintext connector because Credit cases are stored in plaintext file format after DM model evaluation. Plaintext file case base connector is used for persistence of cases. In this connector, the researcher has to specify the path of case structure and also path of text file. All the attributes of a case should be mapped. This is connector's responsibility to retrieve data from case base and return it back to GUI. Like that of case structure, connector is also saved in xml format.



Type: Plain text file

Case base: cbrcase.BasicCBRCaseBase

Case structure file: ba\Desktop\cbe main\caseStructure.xml

Properties

File path: C:\Users\toshlba\Desktop\cbe main\casebase.txt

Delimiter: ,

Mappings

Column	Parameter
0	Description.Current Account Balance
1	Description.sex
2	Description.Marital Status
3	Description.Duration in Current address
4	Description.Occupation
5	Description.No of dependents
6	Description.sucession plan
7	Description.Age
8	Description.Concurrent Credits
9	Description.No of Credits at this Bank
10	Description.Payment Status of Previous Credit
11	Description.Credit Amount
12	Description.Purpose
13	Description.Duration of Credit

Figure 4. 5 Managing Connector Configuration

4.3Manage Tasks and Methods

4.3.1 Managing Tasks

For the development of CBRCADM prototype, the researcher used core package tasks. Core tasks which are used in CBRCADM prototype development are PreCycle, main CBR cycle and PostCycle.

PreCycle task executes before the main CBR cycle. Its task is to get all the cases in case base. Therefore, it is necessary to define path of connector in its subtask. There is only one subtask called obtain case task and it is used retrieve data from case base before the execution of the main CBR cycle.

Main CBR cycle is the main task of CBR cycle and it also has sub tasks. The developer has to give path of case structure in it. It knows number of case attributes that are available. It is called obtain query task. In addition to obtain query task, there are other significant tasks under the main CBR cycle. These are retrieve tasks, reuse tasks, revise task and retain tasks.

Retrieve tasks used to retrieve case(s) from the stored case base. Retrieve tasks also decomposed in to different subtasks. The subtasks include select working cases task, compute similarity task and select the best case. Select working case task selects cases from case base and stores them into current context. Compute similarity task compute similarity of the stored cases with the case entered by the user using the query window. Select best case shows the best matched of case(s) after computing the similarity of stored cases against the new case. It means that the number of best matched case(s) is shown to the user depending on the method used and the threshold.

Reuse tasks enable to reuse previously stored cases. It has three subtasks. These subtasks are: prepare cases for adaptation task, automatic reuse task reuse task. Prepare cases for adaptation task select cases from case base and stores them into context. Here also specifying the path of case structure in this method is needed. Atomic reuse task should be resolved by reuse resolution method.

Revise task is the evaluation stage about the selected solution in reuse phase. After selecting the most similar cases from the retrieved results, the solution for the problem should be confirmed and validated before the solution is stored for future use.

Retain tasks also used to CBR case retention on a persistence layer. It has also its own subtasks like select cases to store task and store cases task. Select cases to store task give authentication to the user for storing case. The store cases task enables to store case(s) into the case base.

The last task in managing tasks in JCOLIBRI is PostCycle. PostCycle task have only one sub task called close connectors task which is usually executed after the main CBR cycle. Its main task is to close a connection between case base and GUI.

Case Similarity, Matching and Ranking

The primarily goal of CBR system is to retrieve best similar cases by using some similarity assessment of heuristic functions. The similarity function involves in computing the similarity between the stored cases in the case base and the query, and selects nearest similar cases to the query. Therefore, JCOLIBRI uses the nearest neighbor algorithm as a cases retrieval technique. This is because JCOLIBRI uses nearest neighbor algorithm for retrieval task. Nearest neighbor algorithm retrieves the case which is nearest to the user`s query by measuring its similarity with the cases. Given a collection of cases and query point in an m-dimensional metric space, find the new case that is closest to the query point. Similar queries are performed by taking a given complex object, approximating it with a high dimensional vector to obtain the query point, and determining the data point closest to it in the underlining feature space.

Nearest neighbor algorithm used to measure the similarity between the stored and the new queries, and return the search results within their ranked order. For each attribute in the query and case, local similarity function measures the similarity between two simple attribute values. Based on the matching weighted sum features from those simple attributes, the similarity score between the queries and stored cases for each simple attribute is assigned.

Finally, the average score (global similarity) of each attribute between the case and the query are computed and the result is assigned to the object (the similarity between the stored case and the query). The maximum degree of similarity among the retrieved cases is displayed according to their ranked order.

4.3.2Managing Methods

The method library stores classes that actually resolve the task. These classes can resolve the CBR cycle using in programming or using GUI. All tasks that are mentioned above should have their own methods to be assigned in order to achieve the tasks goal. The following are lists of methods which are used to solve tasks for this CBRCADM application.

LoadCaseBaseMethod: This method returns the whole available cases from the case base to designer. This method use connector to retrieve case base.

ConfigurQueryMethod: This method resolves obtain query task. By receiving case structure as input parameters, it displays a GUI window so that user can enter query to retrieve cases from the case base.

SelectAllMethod: This method allows displaying all the available cases from the case base to the result window.

SelectSomeMethod: This method resolves to select best task by choosing the ‘n’ most high similarity value from the returned cases. It requests the number of cases to give as input get best match with the requested input.

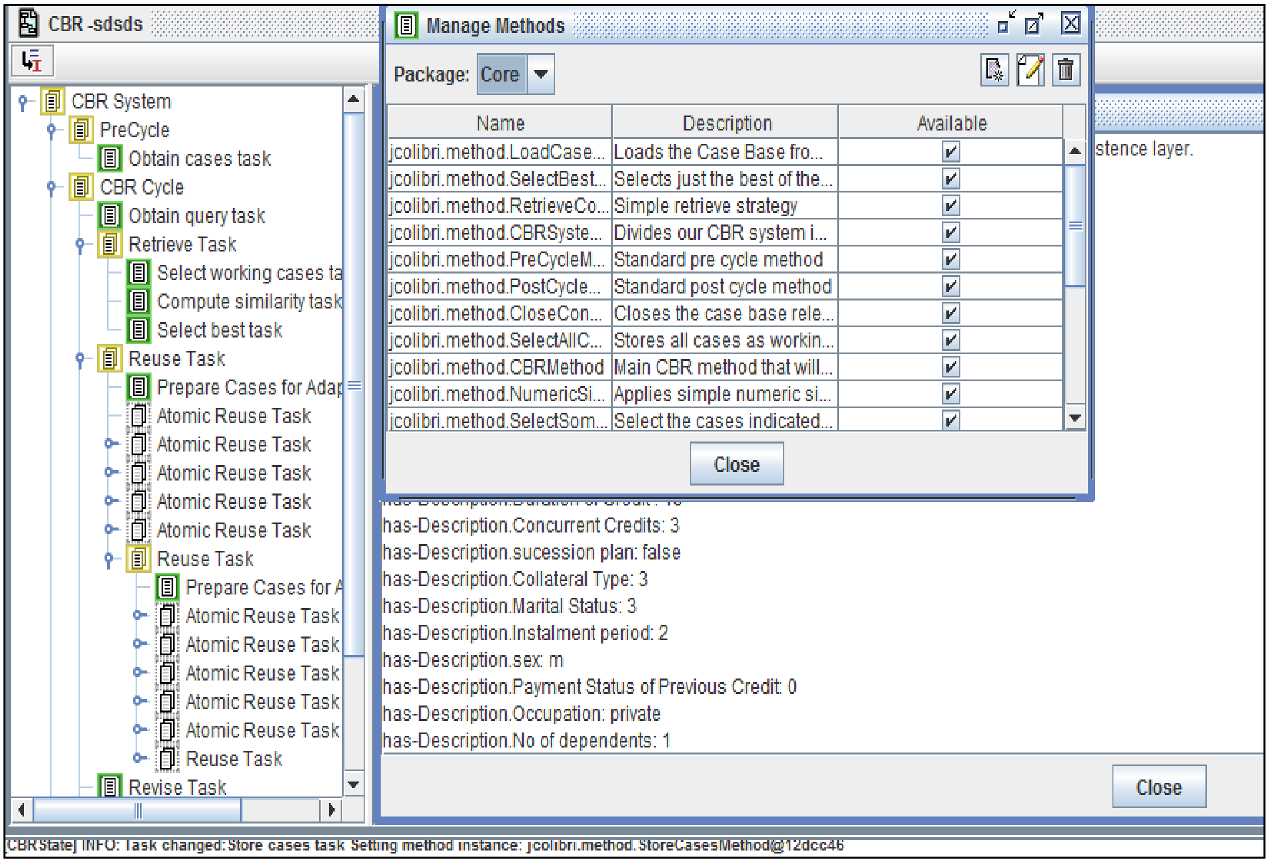
NumeriSimilarityComputationalMethod: this is used to calculate similarity between the query and cases that are stored in the case base.

NumericProportionMethod: it is the sub method of reuse task which involve in computing numeric proportion between the description attributes and solution attributes.

ManualRevisonMethod: Manual revision method permits users to modify cases in the query window as they need.

RetainChooserMethod: This method allows the user to choose the method. Chosen method will store case base. User can choose that he/she want this method to store in case base.

In general, these are some of the methods discussed and used for this research. But, there are many other methods available in JCOLIBRI method library. It is the task of the knowledge engineer to choose the most appropriate method during designing CBR application. Figure 4.6 shows the configuration of tasks and methods. In the configuration window depicted in figure 4.6 the left side shows the tasks and subtask and the right side shows the methods.



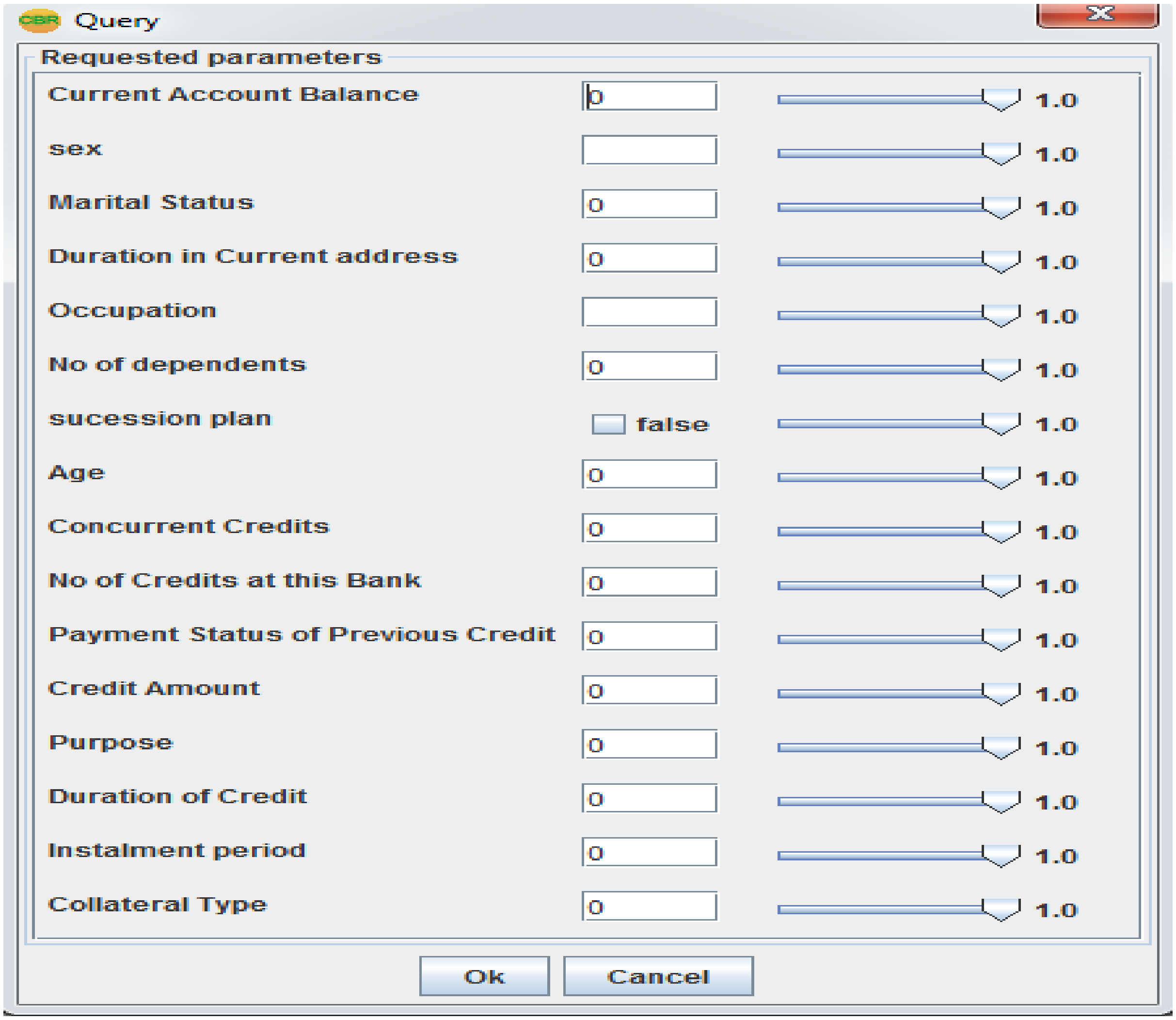
Name	Description	Available
jcolibri.method.LoadCase...	Loads the Case Base fro...	☑
jcolibri.method.SelectBest...	Selects just the best of the...	☑
jcolibri.method.RetrieveCo...	Simple retrieve strategy	☑
jcolibri.method.CBRSyste...	Divides our CBR system i...	☑
jcolibri.method.PreCycleM...	Standard pre cycle method	☑
jcolibri.method.PostCycle...	Standard post cycle method	☑
jcolibri.method.CloseCon...	Closes the case base rele...	☑
jcolibri.method.SelectAllC...	Stores all cases as workin...	☑
jcolibri.method.CBRMethod	Main CBR method that will...	☑
jcolibri.method.NumericSi...	Applies simple numeric si...	☑
jcolibri.method.SelectSom...	Select the cases indicated...	☑

```
has-Description.Concurrent Credits: 3
has-Description.sucession plan: false
has-Description.Collateral Type: 3
has-Description.Marital Status: 3
has-Description.Instalment period: 2
has-Description.sex: m
has-Description.Payment Status of Previous Credit: 0
has-Description.Occupation: private
has-Description.No of dependents: 1
```

```
[CBRState] INFO: Task changed:Store cases task Setting method instance: jcolibri.method.StoreCasesMethod@12dcc46
```

Figure 4. 6 Tasks and Methods Configuration

Retain tasks also used to CBR case retention on a persistence layer. It has also its own subtasks like select cases to store task and store cases task. Select cases to store task give authentication to the user for storing case. The store cases task enables to store case(s) into the case base. The last task in managing tasks in JCOLIBRI is Post Cycle. Post Cycle task have only one sub task called close connectors task which is usually executed after the main CBR cycle. Its main task is to close a connection between case base and GUI. In general, these are some of the methods discussed and used for this research. But, there are many other methods available in JCOLIBRI method library. It is the task of the knowledge engineer to choose the most appropriate method during designing CBR application. Figure 4.6 shows the configuration of tasks and methods. In the configuration window depicted in figure 4.6 the left side shows the tasks and subtask and the right side shows the methods. Figure 4.7 also shows the input or query window of CBRCADM.



Current Account Balance	0	1.0
sex		1.0
Marital Status	0	1.0
Duration in Current address	0	1.0
Occupation		1.0
No of dependents	0	1.0
sucession plan	□ false	1.0
Age	0	1.0
Concurrent Credits	0	1.0
No of Credits at this Bank	0	1.0
Payment Status of Previous Credit	0	1.0
Credit Amount	0	1.0
Purpose	0	1.0
Duration of Credit	0	1.0
Instalment period	0	1.0
Collateral Type	0	1.0

Figure 4. 7 Window for Case Entry into the Case Base

Once the CBR system is ready based on the knowledge acquired using data mining classification technique, the next task is checking its performance. To this end the reseracher evaluate the system performance using test cases and also user acceptance testing.

CHAPTER FIVE
PERFORMANCE EVALUATION OF THE PROTOTYPE

To develop any CBR application, the main building blocks are the previously solved Credit cases which are stored in the case base. Among the main objectives of this research, building the case base is the crucial step for the development of the CBR system. Therefore, to realize this objective, Automatic knowledge acquisition is carried out on CBE Head Office Customers Credit cases to build the case base. For this research a total of 72 instances which scored more than 95 % accuracy during DM test are used to build the case base as discussed in Chapter 3. The evaluation part is basically focused on the performance of the prototype in terms of Precision, Recall and F-measure. In addition the performance of the developed system is evaluated by the potential system users.

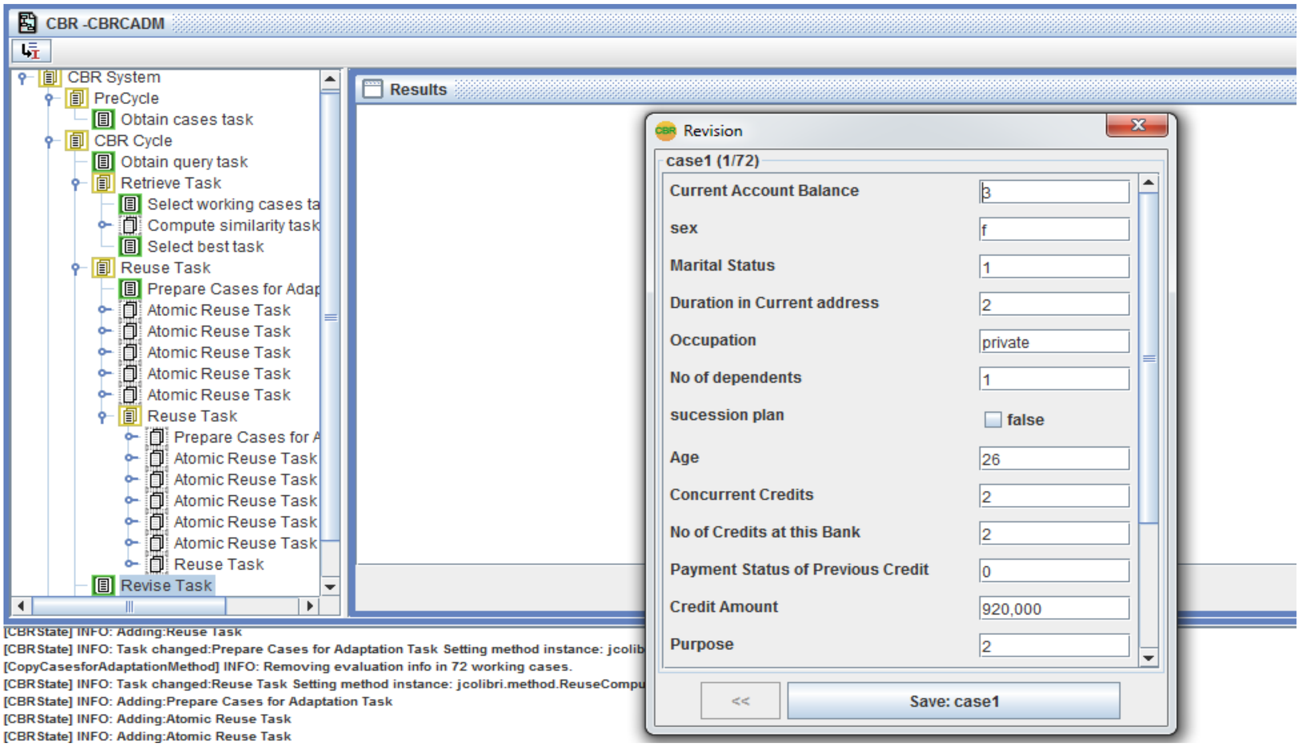
5.1. Testing the CBR Cycles and Evaluating the Performance of the CBRCADM

Now, this is the time to test the functionality of CBR cycles and the soundness of the prototype using selected test cases to check its validity and performance to domain experts. The effectiveness of the prototype is measured with recall, precision and F- measure using test cases. In addition the performance of the system is evaluated from the users' side called user acceptance testing. In this user acceptance testing, potential users' of the system rate the applicability of the system in their day to day activities.

5.1.1. Evaluation of the Retrieval and Reuse Process by Using Statistical Analysis

Retrieval of previously stored cases to solve new problems is the first step in any CBR application. Retrieval of similar cases to the new case from previously solved cases is followed by the reuse of similar solutions. In this research retrieval of cases is performed using the nearest neighbor retrieval algorithm because the implementation tool JCOLIBRI uses this algorithm. During retrieval, similar cases are retrieved to the new case with appropriate ranking. After that the user of the system can use the solution of the retrieved cases in a way that can fit to the problem at hand.

Therefore, retrieval and reuse of cases is successfully implemented in the CBRCADM application as shown in figure 5.1.



case1 (1/72)

Current Account Balance	3
sex	f
Marital Status	1
Duration in Current address	2
Occupation	private
No of dependents	1
sucession plan	□ false
Age	26
Concurrent Credits	2
No of Credits at this Bank	2
Payment Status of Previous Credit	0
Credit Amount	920,000
Purpose	2

```
[CBRState] INFO: Adding:Reuse Task
[CBRState] INFO: Task changed:Prepare Cases for Adaptation Task Setting method instance: jcolib
[CopyCasesforAdaptationMethod] INFO: Removing evaluation info in 72 working cases.
[CBRState] INFO: Task changed:Reuse Task Setting method instance: jcolibri.method.ReuseCompu
[CBRState] INFO: Adding:Prepare Cases for Adaptation Task
[CBRState] INFO: Adding:Atomic Reuse Task
[CBRState] INFO: Adding:Atomic Reuse Task
```

Figure 5. 1 Revise task

The statistical analysis evaluation uses **72** Credit cases that have been collected from CBE Head Office using Automatic Knowledge Acquisition techniques. In this research, the effectiveness of the retrieval process of the CBRCADM is measured by using recall, precision and f- measure. According to McSherry (2001) recall, precision and f- measure are the commonly used measures of performance of the retrieval process in CBR. Recall is the ability of the retrieval system to retrieve all relevant cases to a given new problem (query) from the case base. On the other hand, precision is the proportion of retrieved cases that are relevant to a given query.

To conduct the evaluation, test cases must be prepared and the relevant Credits cases from the case base should be identified. 8 (eight) test cases for testing purposively collected from CBE Jimma District loan department with loan officer’s advice.

Selection of domain experts is done purposively from CBE Jimma District by considering easy of getting the experts and time limitation. The loan Department consisted of 2 loan officers, 2

CRM mangers, 1 credit analyst and 1 loan director. As a result 6 experts are selected for testing the prototype CBRCADM.

For identification of relevant cases, test cases are given to the domain expert in order to assign possible relevant cases from the case base to each of the test cases. The domain expert uses the value of Creditability and Final status (Solution) attributes of the Credit case as the main concept to assign the relevant case to the test cases. After the identification of the relevant cases to the test cases by the domain expert, precision, recall and f- measure are calculated.

The testing method used for evaluating the performance of the prototype system was made by using the parameters precision, recall and F-measure. These three parameters were used in order to measure the accuracy of the prototype system. Recall is defined as the ratio of the number of relevant cases returned to the total number of relevant cases for the new case in case base ; whereas precision is the ratio of the number of relevant cases returned to the total number of cases for a given new cases.(McSherr, 2001). F measure is a derived effectiveness measurement. The resultant value is interpreted as a weighted average of the precision and recall.

Precision and recall can be calculated with the following formulas:

$$Recall = \frac{number\ of\ relevant\ cases\ retrieved}{total\ number\ of\ relevant\ cases}$$

$$Precision = \frac{number\ of\ relevant\ cases\ retrieved}{total\ number\ of\ cases\ retrieved}$$

F-measure is the harmonic mean of precision and recall, with a value between 0 and 1. It is calculated as:

$$F - Measure = \frac{2\ (recall * precision)}{recall + precision}$$

Test Case	Relevant cases from the case base
Case1	case34, case39, case29, case21, case28, case35, case48
Case2	case26, case28, case39, case17, case21, case29
Case 3	case28, case26, case19, case1, case11, case39, case17
Case4	case29, case48, case34, case35, case 63
Case 5	case3, case9, case20, case50, case32
Case 6	case11, case40
Case 7	case15, case17, case19, case21, case23, case65
Case 8	Case3, case 19, case 31,case34

Table 5. 1 Relevant Cases Assigned by the Domain Expert for Sample Test Cases

Once the relevant cases are identified and assigned to the test cases the next step is calculating the recall, precision and f-measure value of the retrieval performance of the CBR system with a threshold interval.

As Henok (2011) indicated in his research, there is no standard threshold for the degree of similarity that has been used for retrieving relevant cases in CBR. Different CBR researchers use different case similarity threshold. Henok (2011) used a threshold level of [1.0, 0.8) i.e. this means cases with global similarity score greater than 80% are retrieved. In this research, the threshold is set by the researcher. For this research, [1.0, 0.8) threshold is used.

Test cases	Relevant cases suggested by domain experts	Relevant cases retrieved by the system	Total cases retrieved by the system	Recall	Precision	F-measure
Test case1	7	6	8	0.85	0.75	0.79
Test case2	6	4	5	0.67	0.80	0.73
Test case3	7	6	9	0.85	0.67	0.75
Test case4	5	4	5	0.80	0.80	0.80
Test case5	6	5	8	0.83	0.62	0.70
Test case6	4	5	7	0.90	0.75	0.82
Test case7	6	6	8	1.0	0.75	0.86
Test case8	4	5	7	0.8	0.71	0.75
Average				0.83	0.73	0.77

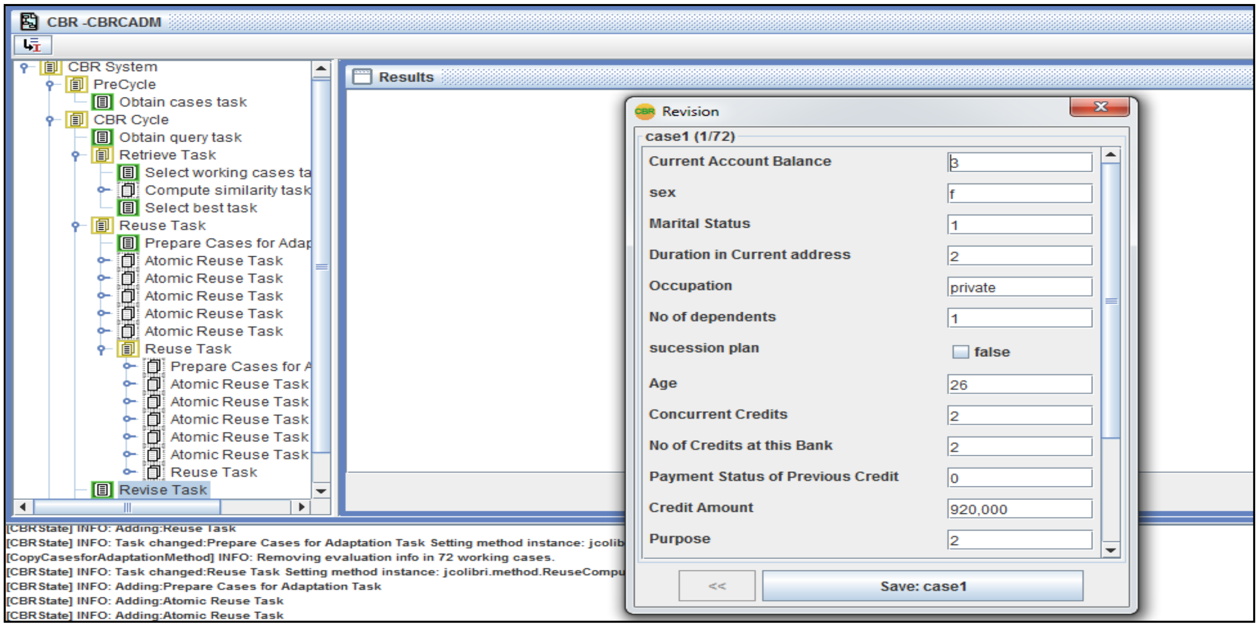
Table 5. 2 Performance Measurement of CBRCADM using Precision, Recall and F-measure

As shown in table 5.2 both recall and precision results are above average which is a hopeful result. The average recall, precision and f- measure results 83%, 73 % and 77 % respectively which is also a promising result. As seen in the table 5.2, for every test case more than average is registered both recall and precision. But, in a complex Banking domain achieving more than average recall and precision is not adequate. 100% or nearer to 100% recall and precision is expected. In terms of recall this research achieved a very good result. But, precision is somewhat lower compared to the average recall. This is because of the tradeoff between precision and recall.

5.1.2 Case Revision and Solution Adaptation Testing

In Financial decision making adaptation is a commonly required task. Since this research main goal is developing credit approval decision making, adaptation is necessary. The purpose testing adaptation of solutions is to evaluate the systems' capability to reuse cases from the case base. Initially the system load case bases at the Pre Cycle stage and then selects working cases from the case base and stores the cases in to current context at the retrieval stage. The next stage is reusing the cases that are loaded in the working memory. If there is no difference between a current case and the retrieved similar cases, null adaptation of solution can be possible. When the previous solution is not fully reasonable in the current problem, only few modifications are required to fit the current situation. This issue is a serious issue especially in credit analysis because of the corresponding risks. Therefore, the adaptation stage requires domain expert knowledge about how differences in problems of previous case and the current situation are occurred. So, it is up to the domain experts to reuse the retrieved cases to solve the new case rather than the system by itself derives solution. Hence, the adaptation stage of CBRCADM is left to the users of the system by comparing specified parameters of the retrieved and current case to modify the solution in a way that can fit to the problem at hand.

In general, the adaptation process of CBRCADM is successful as the case features of the previous and new case have similar or less inconsistency attribute values. On the other hand, no adaptation process can be performed as the attribute values of the previous and new cases have more dissimilar or totally different from the previous cases. However, often a direct application of an uncertain solution is impossible due to the corresponding risks. Therefore, the adaptation has to be performed manually by a human loan expert as shown in the figure 5.2 below.



case1 (1/72)

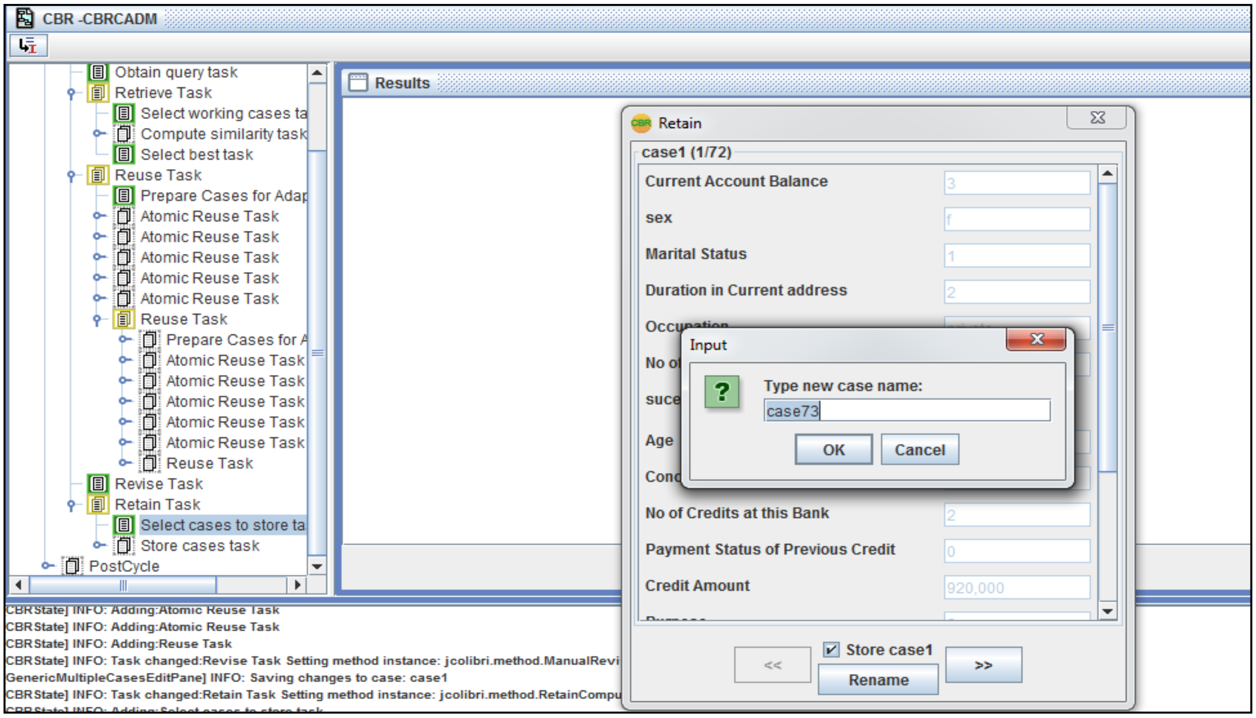
Current Account Balance	3
sex	f
Marital Status	1
Duration in Current address	2
Occupation	private
No of dependents	1
sucession plan	□ false
Age	26
Concurrent Credits	2
No of Credits at this Bank	2
Payment Status of Previous Credit	0
Credit Amount	920,000
Purpose	2

```
[CBRState] INFO: Adding:Reuse Task
[CBRState] INFO: Task changed:Prepare Cases for Adaptation Task Setting method instance: jcolib
[CopyCasesforAdaptationMethod] INFO: Removing evaluation info in 72 working cases.
[CBRState] INFO: Task changed:Reuse Task Setting method instance: jcolibri.method.ReuseCompu
[CBRState] INFO: Adding:Prepare Cases for Adaptation Task
[CBRState] INFO: Adding:Atomic Reuse Task
[CBRState] INFO: Adding:Atomic Reuse Task
```

Figure 5. 2 Revision Interface

5.1.3 Testing Case Retaining

Case retaining is the last cycle in CBRCADM which is an important step in storing new cases which would use for future decision making. Especially, in Credit approval process, retaining cases over time is important because mostly the decision made depends on tacit knowledge and personal experience of the experts. Financial CBR systems should be designed as life-long learning application. CBR systems in bank during loan approval should not be designed only to reuse past episodes with little modification instead retaining new cases for solving similar problems for the future is necessary because rules and guidelines are regularly updated and subjected to change depending on the current situation of the Country. Decisions and target sector change, and new guidelines and procedures emerge constantly. Therefore, the credit decisions for approval, monitoring and Control that were good some years ago may today be obsolete in part. In this research, retaining cases after revision is possible by assigning case index manually as shown in figure 5.3.



case1 (1/72)

Current Account Balance: 3

sex: f

Marital Status: 1

Duration in Current address: 2

Occupation

No of

suce

Age

Cond

No of Credits at this Bank: 2

Payment Status of Previous Credit: 0

Credit Amount: 920,000

Type new case name: case73

☑ Store case1

```
CBRState] INFO: Adding:Atomic Reuse Task
CBRState] INFO: Adding:Atomic Reuse Task
CBRState] INFO: Adding:Reuse Task
CBRState] INFO: Task changed:Revise Task Setting method instance: jcolibri.method.ManualRevi
GenericMultipleCasesEditPane] INFO: Saving changes to case: case1
CBRState] INFO: Task changed:Retain Task Setting method instance: jcolibri.method.RetainCompu
CBRState] INFO: Adding:Select cases to store task
```

Figure 5. 3 Case Retaining Dialog Box

5.2. User Acceptance Testing

User acceptance testing is performed in a real situation at CBE Jimma District with loan officers and CRM Mangers. During testing the users' acceptance, the applicability of the prototype is evaluated by potential users of the system. The loan Department consisted of 2 loan officers, 2 CRM mangers, 1 credit analyst and 1 loan director. Most of them participated in this research work from the beginning to the end by providing the necessary expertise knowledge evaluated the prototype and provide the necessary feedback.

During testing experts are requested to rank each parameter from poor to excellent by assigning value for poor=1, fair=2, good=3, very good=4, excellent= 5.

Evaluation Parameters	**Performance Value**					
	1	2	3	4	5	Average
Adequacy and clarity of result for decision making				4	2	4.3
Relevancy of the retrieved cases in the decision making				4	2	4.3
Fitness of the final solution to the problem at hand				5	1	4.7
Ease of use of the CBR system				6		4
Relevance of the attributes in representing the Credit case			1	3	2	4.7
Efficiency of the system in time			3	2	1	3.7
Resource adequacy of the system		2	2	2		3
Interactivity of the user interface			1	4	1	4
Rate the significance of the system in the domain area				2	4	4.7
					Average	**4.16**

Table 5. 3 Summery of User Acceptance Testing

As depicted in table 5.4, 67% of the respondents' rated CBRCADM as very good and the remaining 33% of the respondents rate as excellent. Similarly, relevance of the retrieved cases in to support users decision making rated very good by 67% of the respondents whereas the remaining 33% of the respondents rate it as excellent. In the case of fitness of the final retrieved solution to the new problem at hand around 83% of the respondents rate the prototype is very good whereas only 17% of the respondents rate as excellent. All the respondents' rate ease of use of the system is 100% very good. 17% of the respondents rate the relevance of attributes in representing credit case rate as very good whereas 50% rate as good and the remaining 33.3% of the respondents' rate as excellent. Lower rating is assigned to the prototype efficiency in terms of time and the resource it has. Only 17% of the respondents' rate the system as excellent in terms efficiency in time. Around 33.3% of the respondents' rates the available resources are fair and in the same way 33.3% and 33.3% of the respondents' rate the resources available are good and very good respectively. The user interface of the prototype is also rated by respondents it is very good around 67% and the remaining 17% is assigned good and 17% excellent. To this end 67% of the respondents rate the applicability of the prototype in their domain excellent and the remaining 37% of the respondents very good. Finally, based on the evaluation of all the respondents the average performance of the prototype is 83.2%, which is above very good. This performance result shows the prototype has a promising applicability for Credit Approval decision making.

In general, the significance of CBR in Credit Approval process has achieved an encouraging result. In addition, relevance and clarity of retrieved cases also rated by the respondents with a highest rating value. One interesting feedback provided by the domain experts is adding explanation facilities of the system. According to the domain experts, explanation facility is important in the adaptation of the retrieved solution. If users of the system get a more explanation about the retrieved solution and the problems itself, user can easily decide whether to use or not use the retrieved solution.

5.3 Discussion

As the researcher discussed in the evaluation section, the proposed system achieved a promising results with system performance of 83% and user acceptance of *83.2%* by using data mining as a main means of knowledge acquisition technique.

The case similarity testing showed that the query is made up of attribute values that have the same value with the case from the case base; the result of the global similarity becomes 1.0. But when there is a difference in the attribute values of the query and the case in the case base, the global similarity value decrease. Therefore adding cases in the case base improve the performance of case based reasoning system in solving problems (new cases). The average recall and precision values for the retrieval performance of the case based reasoning system for Credit approval decision making are 83% and 73% respectively. This indicates that the prototype provides a high percentage of the relevant cases for query which enable the user to give appropriate decision for new credit case. This shows that a promising result is achieved in the study. The reasons that the prototype couldn‵t achieve 100% retrieval and reuse performance could be due to the data and the algorithm used to develop the prototype. Specially attributes quality of financial statement ,risk grade, business outlooks (competitors), industry knowledge and experience, Customer qualification and background, monthly credit reports and account performance which are considered in credit approval are not recorded in the CBE data set and are not included in the case base. This could affect the performance of the prototype. The Nearest Neighbor algorithm, which is used to develop the retrieval process of the prototype, uses distance to compute the similarity between the query and cases by representing the cases in N dimension vector. However the recommendation for the credit application cases doesn‵t have clear boundaries as it has subjectivity and depends on the experience of the domain experts. The performance of the retrieval process and reuse process of the prototype can improve, if all the attributes are included in the research or a way of mechanism that assigns an importance value to the attribute is integrated to the prototype. A hybrid rule based and case based reasoning can be applied for the future. Adding other solved credit cases to the case base can also improve the performance of the prototype.

As per the researcher knowledge, there is no local research attempts made to use CBR for Credit approval decision making, but there are different researches that used CBR or DM for Predicting bank lending decisions. By considering the above performance results of

CBRCADM, it is important to compare with previous CBR System done by Maria & Lusi in 2002in the same area as indicated in the Table 5.5.

Maria & Lusi in 2002 used CBR approach which applied to a survey of bank lending developed by euro system and conducted by the national central bank in each country. The main objective is to enhance the knowledge of credit standards and credit conditions in the euro area. The system uses the data from euro system survey for Portugal. The results obtained lead to the conclusion that the system can forecast with considerable precision (90%) the decision of economic agents.

Title and researcher	Used tool	Case similarity threshold	Performance measurements and results (in %)			Target user
			Recall	Precision	User acceptance	
Predicting bank lending decisions , Maria & Lusi (2002)	Prolog	0.5 – 0.7	Not specified	90%	Not specified	Economic analyst for forecasting the behavior of economic agents in the credit condition
Developing a case based credit approval system using data mining	JCOLIBERI	1.0 - 0.8	83%,	73 %	83.2%	Credit officers, credit analysts and CRM for credit approval decision making

Table 5. 4 Comparison of CBECADM with the previous CBR System work

As shown in Table 5.4 above, Maria & Lusi in 2002 achieved a higher interesting performance in precision (90%) in comparison with this study. The result difference could be due to the increment in attributes that fully express the real working environment and the threshold values difference. Since a more number of cases and attributes can increase the system performance. The current work conducted recall performance measurements and user acceptance testing and registers 83% and 83.3% respectively which are not specified by the former work.

In Maria & Lusi (2002) work the system uses more data from euro system survey for Portugal which makes its own contribution for a better accuracy than the current work which is 83.3 %. Since their work is only considers euro system it can’t be a representative for other areas like our own. As a result the current researches filled this gap by analyzing the existing situations. In addition this the former work attempt to design a tool for the economic analyst for forecasting the behavior of economic agents in the credit conditions and credit standards applied to the approval of loans, whereas the current research attempt to design a system based help for credit experts for credit approval decision process. The current work is also different from Maria & Lusi (2002) since it used the application of data mining for attribute selection and knowledge base construction. But the CBR system developed by Maria & Lusi (2002) used the qualitative data collected through the Bank Lending Survey related to the Portuguese banks.

Besides CBR, different scholars also used data mining techniques for Credit Approval. Credit Approval using Classification Method which aims to evaluate the performance and accuracy of classification models based on decision trees (C5.0 & CART), Support Vector Machine (SVM) and Logistic Regression was performed by Chitra and Subashini in 2013.Since they were not able to obtain a suitable real credit card approval dataset, they used UCI Repository of Machine Learning Databases and Domain Theories. The dataset has 15 attributes plus the class label attribute. They used 690 instances in this dataset, with 307(44.5%) being positive (credit approved) and 383 (55.5%) being negative (credit denied). This research also done with WEKA(Waikato Environment for Knowledge Analysis) which contains a lot of classification algorithms. The results show that the proposed classifiers of CART using J48 algorithm outperform other approaches in solving the problem under investigation with 95.1691 % with 394Correctly Classified Instances of 414 Total Number of Instances. When we compare the outputs Chitra & Subashini (2013) works performs better than the current work but uses a small number of instances. Both researches obtained a better results by using J48 algorithms, they compare different DM algorithms and the results justified that to improve security of the credit approval systems in an automatic and effective way, building an accurate and efficient credit approval system is one of the key tasks for the financial institutions. But the current work is different from the previous research findings with points like:

- ✓ The current work tests the accuracy of the model with test cases and used only selected results for other system as an input.
- ✓ It provides GUI based help for credit case decision makers by using CBR techniques.
- ✓ It uses local dataset for analysis.

In line with this, Choge (2012) conducted a research with the objectives to examine whether naïve Bayes Classifier can be applied accurately to consumer credit evaluation or not. The classification accuracy obtained indicates that the naïve Bayes Classifier has the ability to correctly classify credit applications as either “good” or “bad”. The current work is different from this work since it attempts to use different DM algorithms for model building and got a better result with J48 Classifier and uses the generated knowledge with different CBR framework called JCOLIBERI.

The overall user evaluation for the case based reasoning for credit approval decision making prototype is very good. This shows that the prototype achieved an encouraging result from the perspective of domain experts. The users also suggest the following points for future improvement:

- It is better to include other attributes which are available in hardcopy format to make the system more applicable in the domain.
- It is better to have explanation facilities within the results of the system.
- It is also preferable if the system is web based and accessed from anywhere like other CBE systems.

Based on the above recommendations comes from the users, the researcher tried to direct future research directions and put them as recommendations.

In general, the case based reasoning approach in designing Credit approval decision making system shows an encouraging result for retrieving relevant cases and proposing solution so as to give decisions for new credit cases. It also attain promising user acceptance as it is evaluated by the domain experts. The domain experts (evaluators) assign more than average value for all parameters that are used in the user evaluation form for the prototype. This shows that the prototype achieves an encouraging result from domain expert side in retrieving a ranked order of relevant cases, as well as in proposing a solution to new credit cases. More over the prototype achieved promising result for its speed and easiness to use from the perspective of domain experts.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

Nowadays, the application of AI in banking domain attracts many researchers especially applying the sub field of CBR for risk management. Among the risk that face banks, credit risk is one of great concern to most bank authorities and banking regulators. This is because credit risk is that risk that can easily and most likely prompts bank failure.

Credit risk assessment is very important research field with wide application in the practice. Even if there is a hundreds of research, models and methods, it is still hard to say which model is the best or which classifier or which data mining technique is the best. Each model depends on particular data set or attributes set, so it is very important to develop flexible model which is adaptable to every dataset or attribute set. In order to have better accuracy of model every model should be tested by credit staff because their knowledge can help to improve our models and systems.

The separate application of either data mining or CBR principles cannot fully achieve the aims for evidence based, situation relevant, flexible, and interactive decision making for loan approval. As discussed earlier in previous section, while data mining is dealing with discovering knowledge and model from data, CBR is concerned with how to use that knowledge to solve a new problem. The natural proposition is then that these two approaches can complement each other to better meet the evidence base, situational relevance, flexibility, and better decision making process at the time of approval.

In this study the researcher used a formal approach that combines data mining results and CBR methodologies to provide better decision support for credit approval process of CBE. The rationale for combining data mining and CBR methodologies is to discover knowledge from past data and model using data mining, and to retrieve and enable the use of this knowledge through CBR for the purposes of decision making.

The study was conducted having the main goal of designing and developing a prototype CBR system for credit approval decision making by using manual and automated knowledge acquisition techniques that can assist the domain experts. During the prototype development, real world Credit cases are used from CBE Head office after passing through KDD Model using WEKA 3.6.5 software. Different classification algorithms are tested to build the model and J48 which registers better results is selected for model building. Beside this finding the relevant attributes and cases was carried out by using machine learning algorithms which are available in WEKA. As a results attributes like Purpose, Payment status of previous credit, Collateral Type, Concurrent Credits and Succession plan become the most important features during Credit Approval Decision Making. After the acquired knowledge is modeled, case based reasoning technique is used for representing the knowledge. Cases were represented with attribute-value format. The prototype system CBRCADM is developed by using JCOLIBRI Programming tool.

When measuring the performance of the system, promising results are found. The standard measures of information retrieval (recall, precision and F-measure) are used to measure the retrieval performance of CBR. The average recall, precision and f- measure results 83%, 73 % and 77 % respectively, is also a promising result to apply CBR in the Credit approval decision making. In addition the performance of the system is evaluated by the potential users' of the system and achieved 83.2% performance.

However, the following major critical challenges that need for further investigations were faced while doing this study. The collected credit cases are not enough in size and do not have all important attributes and features used by credit experts during credit approval process. In addition, the general knowledge explanation facility to advise the user when similarity previous solved cases are not found in the case base is not achieved in this study.

6.2. Recommendations

The main goal of this research is to develop a prototype CBR application for Credit Approval Decision Making. At the beginning of this research, the researcher set up different specific objectives in harmony with the overall general objective of this study. To this end, all objectives are achieved successfully with some challenges and constraints. Therefore, there are a number of problems to be investigated by future researchers in applying CBR in credit management.

- CBRCADM has no explanation facility. But, rule-based reasoning has explanation facility and CBR has the capability to incremental learning and specific knowledge acquisition. By combining these interesting capabilities of rule-based reasoning system and CBR system, a hybrid explanation-driven CBR system will improve the performance of CBRCADM.
- The retrieval algorithm used for retrieval of cases for CBRCADM application is nearest neighbor retrieval algorithm. Since the case base of the system increases through incremental learning, the retrieval time increases linearly. Therefore, the retrieval performance will decrease from time to time. In future, there is a need to consider inductive retrieval system that generates a decision tree type structure to organize the cases in memory. In line with these investigating case maintenance techniques is also essential.
- Some of the Credit analysis knowledge such as quality of financial statement, loan officer suggestions about the credit case, relationship analysis, loan amortization documents, commitment letter descriptions, fixed asset reconciliation and Customer follow- up procedures and the like are long texts/documents that cannot be easily converted to rules and cases and they are difficult to include in the rules and/or case-based knowledge based system. Therefore, to incorporate such kind of knowledge in the case base and make it available for potential users, it is better to integrate the case based system with information retrieval.
- Since the current prototype CBRCADM was implemented by using Core JCOLIBERI extensions, it is better to use web interface support extensions by using Tomcat Bridge for the future to have access by all branches and districts like other CBE systems.
- Further research needs to be conducted with the inclusion of other important attributes that have significant impact on credit approval decision making.

REFERENCES

Aamodt, A. & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological

Variations and System Approache, *ALCOM- Artificial intelligence Comunication IOS Press*,7(1), 39-59.

Aamodt, A., Sandtorv,H. and Winnem,O. (1998).Combining Case Based Reasoning and Data

Mining - A way of revealing and reusing RAMS experience, Safety and Reliability*; Proceedings of ESREL, 5(2),1345-1351.*

Abebaw A.(2014). Application of case-based reasoning in legal Case management: an

experiment with Ethiopian Labor law cases, Unpublished, (Masters thesis) Addis Ababa University, Ethiopia.

Alemu, M. (2014). A CBR system for diagnosis and treatment of tubercloses,Unpublished

(Masters thesis). Jimma University,Ethiopia.

Alemu J. (2010): A Case-Based Approach for Designing Knowledge-Based System for AIDS

Resource Center (ARC), Unpublished, (Masters thesis) Addis Ababa University, Ethiopia.

Ali M. & Tickle K. (2009). A Comparison Between Rule Based and Association Rule Mining

Algorithms," in *Third International Conference on Network and System Security*, Gold Coast, pp. 452-455.

Antanassov, A & Antonov, L. (2012). Comparative Analysis of Case-based Reasoning

Frameworks JCOLIBRI and myCBR. Journal *of the University of Chemical Technology and Metallurgy*, 27(1), 83-90.

Antenhe, T (2014), Application of case based recommender system in the tourism sector

for the selection of tourist attraction areas in ethiopia,Unpublished (Masters thesis). ,AddisAbeba University, Ethiopia.

Amritpal S., Amrita K. and ,Jasmeet K.(2015). *Pattern Analysis On Banking Dataset*,

International journal of scientific & technology research ,4(06),23-30.

Ashraf H. & Iqbal, N. (2006). Evaluation of JCOLIBRI: Unpublished Master's Thesis, Maradalen University,Sweden.

Asghar, S. & Iqbal, K. (2009). Automated data mining techniques: *A critical literature review*. *IEEE Proccedings of the International Conference on Information Management and Engineering,* IEEE Xplore Press, Kuala Lumpur, pp: 75-79. DOI: 10.1109/ICIME.

Bank Supervision Directorate. (2010). Bank Risk Management Guidelines, Addisabeba, pp. 01-25: NBE. Retrieved from http://www.nbe.gov.et/pdf/Rm_Guideline revised.pdf

Bank for international settlement (2004, july). Principles for Management of Credit Risk. Retrieved October 15, 2015, from Bank for international settlment website: http://www.bis.org/publ/bcbs108.htm

Bergman, R., Colodner, J., & Plaza, E. (2005). Representation in case-based reasoning. *The Knowledge Engineering Review*, vol 00(0), pp1-4,DOI: 10.1017/S000000000000000.

Bhambri, V. (2011). Application of data mining in banking sector. *International journal of computer science and technology*, 2(2), 199-202. Retrieved 20June,2015,from https://www.academia.edu/1866998/Application_of_Data_Mining_in_Banking_Sector.

Birmingham, W. and Klinker, G. (2009). The Knowledge - Acquisition Tool with Explicit Problem Solving Models. Cambridge Journal , 8(1), 5-25.

Boru, T. (2014). The Determinants of Ethiopian Commercial Banks Performance. *European Journal Of Business And Management*, *6*(14), 65-72.

Buta, P. (1994). Mining for financial knowledge with CBR. AI Expert, 9(2), 34–41.

Buchanan, G. & Forsythe, D. & (1991). Broadening our Approach to Evaluating Medical InformationSystems. *In Proceeding of Annual Symposium on Computer Application in*Medical Care, *pp. 8-12.*

Charlo & Jose, M. (2010). The most relevant variables to support risk analysts for credit decisions. *Regional and Sectoral Economic Studies*, 10(1), 62-70. Retrieved March 22, 2015, from http://www.usc.es/economt/eaa.htm

Chakrabarti .S ,Earl C., Eibe F., Ralf H.G., Jaiwei H. , Xia J., Micheline K., Sam S. L.,Thomas P. ,Richard E. ,Dorian P., Mamdouh R.,Markus S.,Toby J. and Witten H. (2009). *Data mining know it all*. Morgan Kaufmann Publishers 30 Corporate Drive, Suite 400 Burlington, Unite State

Chen, Y.(2009) "Learning classifiers from imbalanced, only positive and unlabeled data sets." Department of Computer Science, Iowa State University.

Charles Y. and Duminda N. (2002), "Modern Intrusion Detection, Data Mining, And Degress Of Attack Guilt," in Applications of Data Mining in Computer Security, pp. 2-25.

Choge J. (2012). Credit evaluation model using naïve bayes classifier: a case of a Kenyan commercial bank, Global *Journal of Computer Science and Technology*, 10(5), 8-17.

Chopra, B., Bhambri , V. & Krishnan, B. (2011). Implementation of data mining techniques for strategic CRM issues. *International. Journal of Computing. Technology*, 2: 879-883.

Chitra K., and Subashini B. (2013), Automatic Credit Approval using Classification Method, International Journal of Scientific & Engineering Research, Volume 4, Issue 7, ISSN 2229-5518.

Collier, K., B. Carey, D. Sautter, C. Marjaniemiand. 1999. *A Methodology for evaluating and selecting data mining software*. Proceedings of the 2nd Hawaii International `Conference on System Sciences, Pp. 1-4.

Compton (1985) Commercial Risk Analysis**,** Canada: John Wiley & Sons, Inc.

Costa, G., Folino,A., Locane, A., Manco, G. & Ortale, R. (2007). Data mining for effective risk analysis in a bank intelligence scenario. *Preccedings of the 23rd International Conference on Data Engineering Workshop,IEEE Xplore Press*, Istanbul, 904-911. DOI: 10.1109/ICDEW.2007.4401083

Cornelius T. (2005), Knowledge Based System Techniques and applications ,1st ed. San Diego, United States of America: Accadamic Press.

Da Silva, I.(2010). Integration of Data Mining and Hybrid Expert System. *In Second European Medical & Biological Engineering Conference FLAIRS.*

Datta R. & Saha S. (2009), An Empirical comparison of rule based classification techniques in medical data bases," in 2nd *International Congress on Pervasive Computing and Management* , Sydney, Australia, pp. 1-15.

Deshpande, M.P. & Thakare, D. (2010). Data mining system and applications: A review. *International Journal of Distributed Parallel Systems*, 1: 32-44.

Dheepa, V. & Dhanapal (2009). Analysis of credit card fraud detection methods, *International Journal of Recent Trends in Enginering*, 2(4): 126-128.

Ethiopia T., (2002). Application of Case-based Reasoning for Amharic Legal Precedent Retrieval: A Case Study with the Ethiopian Labor Law, Unpublished, (Masters thesis) Addis Ababa University,Ethiopia.

Fag, H. & Songdong, J. (2007). Case-Based Reasoning for Logistics Outsourcing Risk Assessment Model. *Proceeding of International Conference on Enterprise and Management Innovation,* pp. 1133-1138.

Ferrario, A. & Smyth, B. (2000). Collaborative Maintenance - A Distributed, Interactive Case-Base Maintenance Strategy. *In Proceedings of the 5th European Workshop on Case-Based Reasoning (EWCBR),* Springer, pp. 393-405.

Henok B. (2011). A Case-Based Reasoning Knowledge Based System for Hypertension Management. Unpublished Master’s Thesis, Addis Ababa University, Ethiopia.

Grimnes, M. & Aamodt, A. (1996). A Two Layer Case-Based Reasoning Architecture for Medical Image Understanding. *In Proceeding of Advances in Case-Based Reasoning*, Berlin: Springer Verlag, pp. 164-178.

Gierl, L. (1993). ICONS: Cognitive basic functions in a case-based consultation system for intensive care. *In the Proceedings of the Artificial Intelligence in Medicine,* AIMED93, pp. 230-236.

Gizaw, M., Kebede, M., & Survalaj, S. (2015). Credit Risk Management and Its Impact onPerformance on Ethiopian commercial Banks. *African Journal Of Business Management*, *9*(2), 59-64.

Goodwin, C. (2005). Research in Psychology: Methods and Design. USA: John Wiley & Sons, Inc.

Hailu, A., & Veni, P. (2015). Credit Risk Management Practice of Ethiopian Commercial Banks. European Journal of Business and Management, 7(7), 01-12. Retrieved February 23, 2015, from http://iiste.org/Journals/index.php/EJBM/article/download/20516/21462

Han, J and Kamber, M (2006), Data Mining: Concepts and Techniques, 2nd edn, Morgan kufman Publishers, San Francisco.

Inderpal S.(2013)" A Review on Knowledge-Based Expert System". International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 2(6), pp1914 1918.

Ionita, I., & Ionita, L. (2011). A decision support based on data mining in e-banking. In *Preccedings of the 10th Reodunet International Conference (RoEduNet)* (p. 5). sydney: IEEE Xplore Press.

Juan A., Antonio A.and Pedro A. (2009). JCOLIBRI 1.0 in a nutshell. A software tool for designing CBR systems, *In Proccedings of the 10th UK Workshop on Case Based Reasoning*, CMS Press,University of Greenwich,

Kazi, I. & Ahmed, Q.(2012) . Use of data mining in banking. Int. J. Eng. Res. Appli.,vol(2)3: 738-742.

Khac, N.A.L. & Kechadi, M.(2010). Application of data mining for anti-money laundering detection: A case study. *Proccedigs of the International Conference on Data Mining Workshop, Dec. 13-13, IEEE Xplore Press, Sydney, NSW.*, pp: 577-584. DOI: 10.1109/ICDMW.2010.66

Lang, S. & Lau, S. (2002). Intelligent Knowledge Acquisition using Case-Based Reasoning. *In the Proceedings of the 13 th Australasian Conference on Information Systems (ACIS2002),* 4-6 Dec., Victory University, Melbourne, Australia. pp. 403-410.

Lawrence J. (1997), *Principles of Managerial Finance*, 11th edition, United States, San Diego State University.

Lenz, M., Bartsch-Sporl, B., Burkhard, H., & Wess, S. (1998). *Case-based Reasoning Technology: From Foundations to Applications.* LNAI: State of the Art. Springer.

Li, W. & Liao, J. (2011). An empirical study on credit scoring model for credit card by using data mining technology. *Proceedigngs of the 7th International Conference on Computational Intelligence and Security*, IEEE Xplor Press, Hainan: 1279-1282. DOI: 10.1109/CIS.2011.283

Maria,T. & Lusi, B. (2002). A case base reasoning approach for predicting bank lending decisions, retrived from http://home.iscte-iul.pt/~luis/papers/ICKEDS06022.pdf, accessed date septembeber 13,2015.

Mantaras, R., Bridge, D., Leake, D., Smyth, B., Craw, S. , Faltings, B., Maher, M., Cox, M., Forbus, K., Keane, M., Aamodt, A. & Watson, I. (2005). Retrieval, reuse, revision, and retention in case-based reasoning: *The Knowledge Engineering Review*, Cambridge University Press, *00* (0), pp 1-2.

McSherry, D. (2001). Precision and Recall in Interactive Case-Based Reasoning. *In Case Based Reasoning Research and Development (ICCBR)*, Lecture notes in Artificial Intelligence, pp. 392-306.

Mekonnen, S. (2009), Credit Risk Management System of Ethiopian Commercial Banks: Case of some public and private banks, Unpublished (Master's thesis), Addis AbebaUniversity,Ethiopia.

Mehdi, M. (2008), Database systems techniques and tools in automatic knowledge acquisition for rule-based expert system, *Knowledge-Based Systems Vol 1*. Washington, United States of America:Acadamic Press, , ch. 8, pp. 201-248.

Mirach, M. (2010), Credit Management: A Case Study of Wegagen Bank Share Company in Tigray Region,Unpublished, Master's thesis in Department of Accounting and Finance College of Business and EconomicsMekelle University, Ethiopia, retrieved on September 16,2015

Mohammed , A. (2013). Towards Integrating Data Mining with Knowledge Based System: The

Case of Network Intrusion detection.Unpublished, M.Sc. Thesis, Addis Ababa University, Ethiopia.

Ngai T., Xiu L. and Chau D. (2009). Application of data mining techniques in customer

relationship management: *A literature review and classification. Expert Syst*. Appli., 36: 2592-2602. DOI: 10.1016/j.eswa.2008.02.021

NBE Directive (2000),Licensing and supervision of banking business directive, Retrieved

from http://www.nbe.gov.et/pdf/Rm Guideline revised.pdf

Pal, S., and Shiu, K. *(2004).Foundation of Soft Cased-Based*Reasoning.Wilely Series on

Intelligent Systems, A John Wiley & Sons, Inc. Publication.

Pulakkazhy, S. & Balan, R. (2013). Data mining in banking and its applications- Review,

Journal of Computer Science , 9 (10): 1252-1259. doi:10.3844/jcssp.

Salem (2005). A Case Base Experts System for Diagnosis of Heart Disease. International Journal

on Artificial Intelligence and Machine Learning, 5(1), pp. 33-39.

Samuel A., Ebenezer F., Gifty A. and Xicang Z. (2012). Risk Factors of Loan Default Payment

in Ghana: A case study of Akuapem Rural Bank:,*International Journal of Academic Research in Accounting, Finance and Management Sciences*,2(4), 22-30.

Schank, C., & Abelson, P. (1977). *Scripts, Plans, Goals and Understanding*. New Jersey,

USA: Erlbaum Hillsdale.

Schumacher, J. & Bergmann, R. (2000). An Efficient Approach to Similarity-Based Retrieval on

Top of Relational Databases. *In Proceedings of the 5th European Workshop on Case-Based Reasoning (EWCBR '2000). Springer.*

Singh,A. Raheja,S. & Kaur,A. (2015). Pattern Analysis On Banking Dataset [Abstract].

International Journal of Scientific & Technology Research, 04(06), 06-21. Retrieved March 21, 2015, from http://www.ijstr.org/final-print/june2015/Pattern-Analysis-On-Banking-Dataset.pdf

Singh, Y., Bhatia, K. & Sangwan, O. (2007). A Review of Studies in Machine Learning Techniques. *International Journal of Computer Science and Security*, 1(1), 70-84.

Thair Nu Phyu (2009), Survey of Classification Techniques in Data mining," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, March 18-20 , pp. 978-988.

Timothy W. (1995) Bank management, 3rd edition United States of America: Harcourt brace publishers.

Triki, S., & Bellamine, N. (2013). Coupling case based reasoning and process mining for a web based crisis Management decision support system. *In Enabling Technologies: Infrastructure forCollaborative Enterprises* Hammame: IEEE, pp. 04-08.

Watson, I. & Marir, F. (1994). Case-based reasoning: A Review. The Knowledge Engineering Review, 9(4), 327-354.

Witten, I & Frank, E (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn, Morgan Kaufmann publishers, San Francisco.

Wilke, W. & Bergmann, R. (1998). Techniques and Knowledge Used for Adaptation During Case-Based Problem Solving. *In Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA '98).*

Yemisrach (2010), Application of case based reasoning in legal knowledge based System, Unpublished, (Masters thesis)   Addis Ababa University, Ethiopia.

Yongjian Fu (2006), Data Mining: Tasks, Techniques and Applications, in *Introduction to Data Mining and its Applications*. Berlin, Germany: Springer Berlin Heidelberg, ch.7, pp. 195-

Zhuang, Z. , Churilov, L., Burstein, F., & Sikaris, K(2009). Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. *European Journal of Operational Research*, 195(3), 662-675.

Appendices

Appendix I: Interview questions to Domain Experts

The main objective of this interview questions is to elicit knowledge from Credit (loan) experts that will help for domain know how and the development of a case-based reasoning system for Credit Approval decision making. The interviewer records the respondents' response using pen, pencil and paper. I thank you in advance for your willingness and valuable time.

1. What is Credit and Credit risk?
2. What criteria's are considered during loan approval process for a particular customer application in Commercial Bank of Ethiopia?
3. What is the basis up on which the bank depends on to pass credit decisions?
4. How do you assess the creditworthiness of a loan applicant?
5. What are the main Credit approval process and procedures that the bank follows to approve credit and which one is the crucial for your decision making process?
6. What are the main risk factors that create credit risk in the bank?
7. What are the main decisions that the loan officers make in credit approval process?
8. What are the most fundamental pre requisites that a customer must fulfill to get credit?
9. What are the major factors that lead the bank to bankruptcy in credit management? And how can you manage the challenges you face?

Appendix II: Prototype Evaluation form for the Domain Expert

This is an evaluation form to be filled by loan experts in order to evaluate the applicability of the case-based reasoning system in Credit Approval Decision Making. I thank you in advance for your willingness and valuable time.

Description of the parameter values are as follows.

Performance Value	1	2	3	4	5
Description	Poor	Fair	Good	Very good	Excellent

Instruction: Please, tick on the appropriate value for the corresponding parameter of the case-based reasoning system in Credit Approval Decision Making.

Evaluation Parameters	**Performance Value**					
	1	2	3	4	5	Average
Adequacy and clarity of result for decision making						
Relevancy of the retrieved cases in the decision making						
Fitness of the final solution to the problem at hand						
Ease of use of the CBR system						
Relevance of the attributes in representing the Credit case						
Efficiency of the system in time						
Resource adequacy of the system						
Interactivity of the user interface						
Rate the significance of the system in the domain area						
			Average			

Appendix III: Dataset Description

No	Attributes	Description	Values	Type
1	Current account balance	Current account balance of the customer in the Bank	No running account 1<=…<100,000 birr More than100,000 birr	Numeric
2	Duration in current address	Duration in current address of the customer in years	Unknown, <=1year, 1-4 years,	Numeric
3	Age	Age of the customer	Numeric values	Numeric
4	Sex	Gender characteristics of the customer	f- female	Nominal
5	Marital status	Marital status of customer	Single Divorced	Numeric
6	Concurrent credits	Further running credits	At this bank	Nominal
7	Occupation	Job type of the customer	Pi-private	Nominal
8	No of credits at this Bank	Total number of credit at this Bank (this shows the relationship of the customer with the bank)	Numeric value	Numeric
9	No of dependents	Number of persons under his full support and control	Numeric value	Numeric
10	Payment status of previous credit	Customers credit re payment status	No previous credit current	Nominal
11	Credit amount	The amount of the credit in birr	Numeric value	Numeric
12	Purpose	Purpose of the credit (reason for the credit request)	Working capital, fixed asset, domestic trade and services, agriculture, manufacturing, import	Nominal
13	Duration of credit	Duration of credit in months	Numeric value	Numeric
14	Installment period	Installment period of the credit	Monthly, Quarterly	Numeric
15	Collateral type	The collateral the customer provides as a guarantor for the credit	Building/house, vehicle, Machinery, Bond, Farm, Other	nominal
16	Succession plan	The succession plan readiness of the customer	Yes or No	Nominal
17	Creditability	Whether the customer is credible or not	Yes or No	Nominal
18	Final status	Repayment status of the Customer	Approved good, Non-performing or Denied	Nominal

Appendix IV: **J48 CLASSIFIER OUTPUTS**

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: CBE main dataset22-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R12

Instances:1518

Attributes:17

            Current Account Balance

            sex

             Marital Status

            Duration in Current address

            Occupation

            No of dependents

            sucession plan

            Age

            Concurrent Credits

            No of Credits at this Bank

            Payment Status of Previous Credit

            purpose

            Duration of Credit

            Instalment period
```

```
              Collateral Type

              Creditability

              Final status

Test mode:split 90.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree

------------------

succession plan = yes

|   No of dependents <= 2

|   |   Current Account Balance <= 2

|   |   |   Concurrent Credits = No further running credits: denied (158.3/70.43)

|   |   |   Concurrent Credits = At this bank

|   |   |   |   purpose = fixed asset

|   |   |   |   Marital Status =  married

|   |   |   |   |   Collateral Type <= 2: approved good (6.0)

|   |   |   |   |   Collateral Type > 2

|   |   |   |   |   |    Marital Status =  Single : denied (0.0)

|   |   |   |   |   |    Marital Status = divorced: denied (4.0)

|   |   |   |   |   |    Marital Status = single: denied (3.0/1.0)

|   |   |   |   |   |    Marital Status = 4.0: denied (0.0)

|   |   |   |   purpose = Manufcaturing: denied (1.0)
|  |  |  |  purpose = Domestic trade
|  |  |  |  |  Payment Status of Previous Credit = special mention: denied (3.0)
|  |  |  |  |  Payment Status of Previous Credit = no previous credit: approved good (6.0/1.0)
|  |  |  |  |  Payment Status of Previous Credit = spacial mention: approved good (7.0/1.0)
|  |  |  |  |  Payment Status of Previous Credit = sub standard: approved good (3.0/1.0)
|  |  |  |  |  Payment Status of Previous Credit = doughtfull: denied  (0.0)
|  |  |  |  purpose = Agri: approved good (22.0/2.0)
|  |  |  |  purpose = working cap: approved good (7.0)
|  |  |  |  purpose = domesic trade
|  |  |  |  |  Current Account Balance <= 1: denied (10.0)
|  |  |  |  |  Current Account Balance > 1
|  |  |  |  |  |  Instalment period = yearly: Denied (3.0/1.0)
|  |  |  |  |  |  Instalment period = quarterly: approved good (3.0)
|  |  |  |  |  |  Instalment period = semianually: approved good (2.0)
|  |  |  |  |  |  Instalment period = monthly: approved good (0.0)
|  |  |  |  |  |  Instalment period = 4.0: approved good (0.0)
|  |  |  |  purpose = working cap or domesic trade
|  |  |  |  |  Duration in Current address <= 3: denied (2.0)
|  |  |  |  |  Duration in Current address > 3: approved good (3.0/1.0)
|  |  |  |  purpose = import/export: approved good (0.0)
|  |  |  Concurrent Credits = At other banks: approved good (190.0/27.0)
|  |  Current Account Balance > 2: approved good (568.7/81.13)
|  No of dependents > 2: approved good (216.0/7.0)
sucession plan = no
|  purpose = fixed asset: non performing (53.0/31.0)
|  purpose = Manufcaturing: denied (3.0)
|  purpose = Domestic trade
|  |  Duration of Credit  <= 15: approved good (49.0/27.0)
|  |  Duration of Credit  > 15: non performing (53.0/30.0)
|  purpose = Agri: non performing (20.0/10.0)
|  purpose = working cap: non performing (33.0/17.0)
|  purpose = domesic trade: denied (69.0/29.0)
|  purpose = working capdomesic trade: denied (1.0)
|  purpose = import/export: denied (15.0/6.0)
Creditability = no: denied (315.0/5.0)
Number of Leaves  :   38
Size of the tree :        59
Time taken to build model: 0.08seconds
=== Evaluation on training split ===
=== Summary ===
Correctly Classified Instances       1233               90.2635 %
Incorrectly Classified Instances      133                9.7365 %
```

```
Kappa statistic                      0.7587

Mean absolute error                       0.1145

Root mean squared error                      0.24

Relative absolute error                 37.5457 %

Root relative squared error             61.3574 %

Total Number of Instances              1366

=== Detailed Accuracy By Class ===

             TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class

               0.996     0.312      0.881     0.996     0.935      0.842    approved good

               0         0          0         0         0          0.611    non performing

               0.996     0.005      0.983     0.996     0.989      0.996    denied

Weighted Avg.    0.903     0.219      0.82      0.903     0.859      0.852

=== Confusion Matrix ===

   a   b   c   <-- classified as

 952   0   4 |   a = approved good

 127   0   1 |   b = non performing

   1   0 281 |   c = denied
```